# Rethinking the Intelligence Cycle: Defensive Hybrid Intelligence as an Integrative Theory for the AI Era

Çağdaş İNAL

ORCID: https://orcid.org/0009-0006-2721-8749 Email: cagdasinal80@gmail.com

Article Type: Conceptual research article Status: Working draft, July 2026

Author Note:

This conceptual framework is developed as part of the foundational research for the forthcoming book *İstihbarat Teorisi 2026*. The author declares no conflict of interest.

# Contents

## Abstract

**Background:** Established models of intelligence production — the intelligence cycle, the target-centric model and the OODA loop — were conceived for an environment in which collection was scarce, analysis was human, and adversarial manipulation of the information space was peripheral rather than central. The convergence of large language models, ubiquitous open-source data and industrialised influence operations has inverted each of these assumptions.

**Purpose:** This article develops Intelligence Theory 2026, an original conceptual framework organised around the construct of Defensive Hybrid Intelligence (DHI). The framework integrates artificial intelligence, open-source intelligence, information fusion, explainable AI, human–machine teaming and intelligence governance into a single theoretical architecture for strategic decision-making under adversarial information conditions.

**Methods:** The study employs qualitative conceptual research combining theory synthesis, a concept-centric structured review of ten thematic literatures, comparative theoretical synthesis and formal framework construction. Constructs are defined, causal relationships specified, and testable propositions derived.

**Results:** The framework identifies four interacting layers — sensing, fusion, judgement and governance — connected by explicit feedback loops, and repositions cognitive defence from an afterthought to a structural property of the intelligence enterprise. Six propositions specify the conditions under which hybrid human–AI configurations outperform either fully automated or fully human analytic arrangements.

**Contribution:** The article extends intelligence theory beyond process description towards an integrative, governance-aware model that treats epistemic security as a first-order design requirement.

**Implications:** The framework offers agencies, corporate intelligence functions and AI developers a common vocabulary for designing accountable, explainable and manipulation-resistant intelligence systems.



**Keywords:** *intelligence theory; defensive hybrid intelligence; open-source intelligence; artificial intelligence; explainable AI; information fusion; human–machine teaming; cognitive warfare; intelligence governance; strategic decision-making*

# Research Questions, Objectives and Propositions

## Research questions

RQ1. In what respects do canonical intelligence models — the intelligence cycle, target-centric analysis and the OODA loop — fail to accommodate AI-enabled collection, analysis and adversarial manipulation?

RQ2. What constructs are necessary to theorise intelligence production in which human analysts and machine systems share analytic labour?

RQ3. How should open-source intelligence be positioned theoretically when it constitutes the dominant, rather than supplementary, collection stream?

RQ4. Under what conditions does explainability function as a precondition for analytic trust rather than as a technical property of models?

RQ5. How can cognitive defence against foreign information manipulation and interference (FIMI) be integrated into intelligence theory as a structural layer rather than as a separate discipline?

RQ6. What governance mechanisms are required to preserve accountability when analytic judgements are co-produced by humans and AI systems?

RQ7. Through what feedback mechanisms does an AI-enabled intelligence architecture learn, correct error and resist degradation over time?

## Research objectives

RO1. To evaluate critically the theoretical adequacy of existing intelligence models against the technological and adversarial conditions of the mid-2020s.

RO2. To construct and formally define Defensive Hybrid Intelligence as the core construct of an integrative framework, Intelligence Theory 2026.

RO3. To specify the framework's components, layers, causal relationships and feedback loops with sufficient precision to permit future empirical testing.

RO4. To compare the framework systematically with the intelligence cycle, the DIKW hierarchy, the OODA loop and hybrid intelligence models, identifying advantages and limitations.

RO5. To derive governance and design implications for state, corporate and academic intelligence practice.

## Theoretical propositions

P1. Holding analytic resources constant, intelligence architectures that structurally separate machine-scale sensing from human-scale judgement, with an explicit fusion layer between them, will exhibit lower rates of unrecognised assessment error than architectures that

collapse these functions; the advantage increases with the adversarial intensity of the information environment (moderator: adversarial intensity).

P2. The analytic value of machine-generated output is conditional on decomposability: as decomposability falls, analyst reliance decouples from actual system reliability, manifesting as automation bias where salient system errors have not been observed and as algorithm aversion where they have (mediator: trust calibration; moderator: error salience).

P3. In information environments subject to sustained FIMI activity, intelligence systems lacking a dedicated cognitive-defence layer will exhibit progressive corruption of their open-source input streams; the corruption rate increases with the share of machine-mediated ingestion and decreases with the coverage of immutable provenance capture (moderators: machine-mediation share; provenance coverage).

P4. Human–AI teaming outperforms both fully human and fully automated analysis when task allocation follows comparative epistemic advantage assessed under adversarial conditions: pattern detection and volume processing to machines, contextual judgement and adversarial reasoning to humans; the performance gap widens with task equivocality and with adversarial injection into the task environment (moderators: equivocality; injection intensity).

P5. Governance mechanisms embedded at design time — auditability, provenance tracking, model documentation — reduce undetected intelligence-failure rates more effectively than post-hoc oversight applied to completed assessments; the effect strengthens with the proportion of the enterprise's consequential choices made at design time (moderator: design-time decision share).

P6. Feedback-loop quality (L1–L4) determines an intelligence system's error-correction capacity over repeated cycles more strongly than the sophistication of any single component; the dominance of loop quality increases with environmental volatility and with the adversary's adaptation rate (moderators: volatility; adversary adaptation rate).

# 1. Introduction

The theoretical vocabulary of intelligence studies remains, to a striking degree, the vocabulary of the mid-twentieth century. The intelligence cycle — requirements, collection, processing, analysis, dissemination — entered doctrine in an era when collection was expensive, secret and slow; when analysis was the exclusive province of human specialists; and when deliberate manipulation of the open information environment was a peripheral counter-intelligence concern rather than a structural condition of the environment itself. Kent's (1949) account of strategic intelligence as organised knowledge in the service of statecraft, and Kendall's (1949) contemporaneous critique of it, framed a debate about the relationship between analysis and policy that the discipline has refined but never fundamentally reconstructed. Later scholarship questioned the cycle's descriptive accuracy (Hulnick, 2006), proposed target-centric alternatives (Clark, 2019) and interrogated the epistemic status of intelligence knowledge (Rønn & Høffding, 2013). Yet the underlying architecture — discrete functions, broadly sequential flow, human judgement at the centre — has proved remarkably durable.

Three assumptions embedded in that architecture have now been inverted. The first concerns scarcity. Classical theory treats information as a scarce commodity extracted from denied environments at cost and risk; the contemporary condition is one of abundance, in which the binding constraint has shifted from access to attention, triage and verification (Zegart, 2022). The second concerns the locus of analysis. Where analysis was once an irreducibly human activity supported at the margins by computation, machine systems now perform substantive analytic work — entity extraction, translation, pattern recognition, summarisation, first-draft assessment — at a scale no human workforce can match (Vogel et al., 2021). The third concerns the integrity of the information environment. Foreign information manipulation and interference is no longer an occasional nuisance at the boundary of the collection problem; it is an industrialised, continuous activity that targets the very open sources on which modern intelligence increasingly depends (Goldstein et al., 2023). A theory built on scarcity, human analytic monopoly and a broadly honest information environment cannot simply be patched to accommodate the negation of all three premises.

The scholarly response to this transformation has been substantial but fragmented. Students of intelligence have examined artificial intelligence chiefly as a tool to be absorbed into existing processes, asking how machine learning might accelerate particular stages of the cycle (Vogel et al., 2021; Zegart, 2022). Scholars of strategy have debated whether prediction technologies substitute for or complement human judgement in war (Goldfarb & Lindsay, 2022; Payne, 2021). The OSINT literature has documented the discipline's growth from marginal supplement to primary collection stream without resolving its theoretical position within the wider enterprise (Glassman & Kang, 2012; Williams & Blum, 2018). Research on explainable AI has matured largely inside computer science, with limited traffic across the disciplinary boundary into intelligence studies (Miller, 2019; Vogel et al., 2021). Work on disinformation and cognitive security has, for its part, developed its own frameworks — notably the DISARM taxonomy and the FIMI analytic tradition — in near isolation from

intelligence theory proper. Each of these literatures illuminates one face of the problem. None integrates them.

The arrival of large language models has sharpened the problem rather than merely extended it. Foundation models trained on internet-scale corpora (Bommasani et al., 2021; Brown et al., 2020) collapse the boundary between collection and analysis: the same system that retrieves open-source material can summarise, translate, correlate and characterise it in a single pass, in prose indistinguishable from an analyst's first draft. The efficiency gain is obvious. Less obvious, and less examined, are its theoretical consequences. Machine-generated analytic products inherit the opacities of the models that produce them; their errors are systematic rather than random; and the corpora from which they learn are precisely the open sources that adversaries have the strongest incentive to poison. An intelligence enterprise that adopts such systems without a theory of how machine judgement, human judgement and adversarial pressure interact is not modernising its doctrine — it is outsourcing its epistemology to systems it cannot interrogate.

This is the gap the present article addresses. What is missing from the literature is not another taxonomy of AI applications, nor a further critique of the intelligence cycle, but an integrative theoretical framework that treats machine analysis, open-source dominance and adversarial information conditions as jointly constitutive of the intelligence problem, rather than as three separate complications of it. Because the phenomenon in question is a configuration of concepts rather than a body of data awaiting measurement, the appropriate mode of inquiry is conceptual: systematic engagement with the relevant literatures, comparative synthesis of competing models, and the construction of a framework whose constructs are defined precisely enough to be tested by others. The methodology section defends this choice at length; here it suffices to note that intelligence studies has a long tradition of theory-building of exactly this kind, from Kent (1949) through Herman (1996) to Omand (2010), and that the discipline's present difficulty is theoretical before it is empirical.

The article's contribution is a framework we term Intelligence Theory 2026, organised around the construct of Defensive Hybrid Intelligence (DHI). The framework models the intelligence enterprise as four interacting layers — sensing, fusion, judgement and governance — connected by explicit feedback loops, and makes two moves that distinguish it from its predecessors. First, it treats the human–machine division of analytic labour as a design variable to be theorised, not an implementation detail: the framework specifies where machine processing ends, where human judgement begins, and what the interface between them must guarantee (above all, explainability) for the whole to remain trustworthy. Second, it repositions cognitive defence — the protection of the enterprise's own inputs, models and analysts from adversarial manipulation — as a structural layer of intelligence production rather than an adjacent speciality. From these moves the framework derives six testable propositions concerning the conditions under which hybrid configurations outperform fully human or fully automated alternatives.

The argument unfolds in three main movements. First, I examine the theoretical debt we owe to classical models and where those models abruptly fail under modern adversarial conditions. Section 2 reviews the relevant literatures; intelligence theory, strategic

intelligence, OSINT, artificial intelligence, decision analytics, information fusion, cognition and bias, explainability, ethics, human–machine teaming and governance — identifying in each the disagreements and limitations that motivate the framework. Section 3 consolidates these into a statement of the theoretical gap. Section 4 develops the framework itself: constructs, definitions, causal relationships, feedback loops and decision architecture. Section 5 sets out the methodology. Section 6 compares the framework with the intelligence cycle, the DIKW hierarchy, the OODA loop and hybrid intelligence models. Sections 7 to 10 address theoretical contributions, practical implications, limitations and future research, before Section 11 concludes.

# 2. Literature Review

The review that follows is organised thematically rather than chronologically. Each subsection identifies the principal positions in a literature, the disagreements between them, the theoretical limitations they share and the questions they leave unresolved. The purpose is not encyclopaedic coverage but the assembly of the specific materials from which Section 3 constructs the gap and Section 4 constructs the framework. The strands are taken in turn across subsections 2.1 to 2.12.

## 2.1 The evolution of intelligence theory

Intelligence studies has long been described as a field rich in history and poor in theory. Kent (1949) supplied the founding statement: intelligence is knowledge, organisation and activity, and its purpose is to inform policy while remaining scrupulously separate from it. Kendall (1949), reviewing Kent, objected at once that this separation was neither achievable nor desirable, and that analysis divorced from the policymaker's actual questions would be ignored. The Kent–Kendall exchange established a fault line — proximity versus distance between analysis and policy — that runs through the discipline to the present day and that, as will become clear, acquires new force when the 'analyst' in question is partly a machine.

A second fault line concerns definition itself. Warner (2002) demonstrated that the field lacked even an agreed definition of its object, and proposed secrecy as the differentia: intelligence is secret state activity directed at foreign entities. That definition sat uneasily with practice even in 2002 and sits far more uneasily now, when the majority of exploitable material is openly available and much intelligence work is performed by non-state actors. Herman (1996) had earlier offered a more capacious institutional account — intelligence as a form of state power resting on organised information — while Gill and Phythian (2018) broadened the object further, treating intelligence as the mediation of security threats through surveillance, analysis and action by a range of actors. The disagreement is not merely semantic. A secrecy-centred definition marginalises OSINT by construction; an information-power definition accommodates it but says little about how it should be organised.

The third and most consequential debate concerns process. The intelligence cycle became the canonical process model largely through pedagogy and doctrine rather than through theoretical defence (Lowenthal, 2022). Hulnick (2006) mounted the best-known attack, showing that the cycle misdescribes practice at almost every stage: collection and analysis proceed in parallel rather than in sequence, policymakers rarely originate requirements, and dissemination feeds back into collection continuously. Clark (2019) responded constructively with the target-centric model, replacing the linear cycle with a collaboratively maintained model of the target to which collectors, analysts and consumers all contribute. Phythian (2013) assembled the fullest critical examination, in which several contributors concluded that the cycle survives because it is administratively convenient, not because it is true. Treverton (2009) added the environmental argument: the shift from puzzles (questions with discoverable answers) to mysteries (contingent futures) undermines the cycle's implicit assumption that intelligence problems have collectable solutions.

A fourth strand interrogates the epistemology of the enterprise. Betts (1978) argued that intelligence failure is inevitable in the strict sense: the pathologies that produce failure are structural features of the relationship between analysis and decision, not removable defects. Rønn and Høffding (2013) pressed the epistemological question directly, asking what kind of knowledge intelligence claims to be and concluding that it is best understood as a species of judgement under uncertainty rather than as descriptive knowledge simpliciter. Marrin (2017) has argued across a series of interventions that the field's improvement depends on systematic borrowing from other disciplines — a prescription this article follows. What none of these literatures anticipated is a condition in which part of the judging is done by statistical systems whose 'epistemology' — training data, objective functions, failure modes — differs in kind from human cognition. The evolution of intelligence theory has, in short, produced sophisticated critiques of the classical model and no successor of comparable generality.

## 2.2 Strategic intelligence, military intelligence and national security

The strategic intelligence literature addresses the enterprise's highest-level purpose: warning, estimation and the support of state decision-making. Omand (2010) provides the most systematic modern statement from a practitioner-scholar, framing intelligence as the reduction of decision-makers' ignorance within an explicitly ethical and legal envelope, and proposing elucidation, explanation and prediction as its ascending functions. Gentry and Gordon (2019) focus the lens on strategic warning, documenting how warning as a distinct discipline has atrophied inside large agencies even as the demand for it has grown. Dahl (2013) complicates the warning problem from the consumer's side: his study of surprise attacks finds that strategic-level warning is chronically available and chronically unpersuasive, whereas decision-makers act on specific tactical warning — a finding with sharp implications for any framework promising AI-enhanced foresight, since better strategic prediction may simply produce more elegant reports that go unheeded.

Within military intelligence, the debate has turned on what algorithmic systems change. Horowitz (2018) treats AI as a general-purpose technology whose military significance lies in diffusion and organisational adaptation rather than in any single capability. Payne (2021) argues that machine tactical acumen coexists with machine strategic incompetence: systems that excel at bounded optimisation lack the theory of mind that strategy requires. Goldfarb and Lindsay (2022) generalise the point economically — machine prediction is a complement to, not a substitute for, human judgement, and cheaper prediction raises the value and the burden of the judgement and data infrastructure surrounding it. Jensen, Whyte and Cuomo (2020) supply the organisational corrective: the limits of military AI lie less in the algorithms than in the institutions that must absorb them, with their attendant bureaucratic politics, brittleness and new error surfaces. Lahneman (2010) had already argued that the transformation under way amounts to a paradigm change requiring intelligence to fuse open and secret knowledge streams as co-equals; Davies (2004) reminds us that such adaptation is refracted through national intelligence cultures that differ profoundly even between close allies.

Zegart (2022) draws these threads together in the most influential recent synthesis: the erosion of the state's collection monopoly, the rise of open-source competitors from commercial imagery firms to volunteer investigators, and the widening gap between the speed of technological change and the adaptive capacity of intelligence bureaucracies. The synthesis is diagnostic, however, and deliberately so. The strategic literature thus converges on a shared conclusion — that AI and open information transform the conditions of strategic intelligence — while diverging on mechanism (diffusion, complementarity, organisational absorption) and offering no integrated model of how a strategic intelligence enterprise should be architected under the new conditions. That absence is precisely the space in which the present framework operates.

## 2.3 Open-source intelligence

OSINT scholarship has passed through three broadly identifiable phases. In the first, advocacy, practitioner-scholars argued for the discipline's admission to the collection canon. Steele (2007) made the maximalist case — that the overwhelming majority of what decision-makers need is openly available and that secret collection should be the exception — while Mercado (2004) argued from within the profession that open sources routinely outperform clandestine ones on speed and breadth. The advocacy literature succeeded rhetorically but left OSINT theorised as a cheaper substitute for secret collection: same epistemology, different price.

The second phase treated OSINT as transformative rather than substitutive. Glassman and Kang (2012) argued that internet-era OSINT is not a collection discipline at all in the classical sense but a different mode of problem-solving, distributed and iterative, that changes what intelligence is rather than merely how much of it there is. Williams and Blum (2018) formalised the periodisation, distinguishing a document-retrieval first generation from a second generation defined by social media exploitation, big-data analytics and the erosion of the boundary between collection and analysis. Akhgar, Bayerl and Sampson (2016) consolidated the investigative-methods strand, particularly for law enforcement, while Pastor-Galindo et al. (2020) mapped the computational research agenda — crawling, entity resolution, network analysis, credibility assessment — and catalogued its unsolved problems, verification chief among them.

The third phase is critical and governance-oriented. Hribar, Podbregar and Ivanuša (2014) identified the legal and ethical grey zone in which OSINT operates: 'open' does not mean 'ethically unencumbered', and the aggregation of individually innocuous public data can constitute surveillance in all but name. Ghioni, Taddeo and Floridi (2024), reviewing the governance and ethical literature systematically, find it thin, fragmented and lagging far behind capability — a finding that should unsettle a discipline whose raw material is other people's public lives. Meanwhile the practical centre of gravity has shifted decisively: commercial satellite imagery, open flight and shipping data and volunteer analytical communities now produce assessments of strategic consequence in public, in near real time (Zegart, 2022). Three questions remain unresolved across all phases. First, verification: no scalable theory exists for establishing the reliability of open sources under adversarial conditions, and the sources most freely available are the easiest to poison. Second, position:

whether OSINT is a collection discipline among others or the substrate of the entire enterprise remains contested. Third, epistemic exposure: the literature has barely begun to address what it means that the corpus OSINT exploits is also the corpus on which large language models are trained and against which influence operations are waged. The present framework takes a definite position on all three.

## 2.4 Artificial intelligence, machine learning and large language models

The technical trajectory relevant to intelligence can be stated briefly. The transformer architecture (Vaswani et al., 2017) made it feasible to train very large models on very large corpora; Brown et al. (2020) demonstrated that scale alone yields few-shot competence across tasks for which the model was never explicitly trained; and Bommasani et al. (2021) named the resulting class — foundation models — while cataloguing its defining property: capabilities and defects alike are inherited downstream by every application built on the model, most of them poorly understood at the point of inheritance. For intelligence work the significance is double-edged. A single system can now read, translate, summarise, correlate and draft across the entire open-source corpus; and the same system carries into every product it touches the biases of its training data, the opacity of its internal reasoning and a documented propensity to generate fluent falsehood.

The intelligence-studies reception of these developments divides into identifiable camps. An augmentation school, represented by Vogel et al. (2021) and by much of Zegart (2022), holds that AI's proper role is the automation of analytic drudgery — triage, translation, entity extraction — freeing human analysts for judgement, and identifies collaboration, transparency and accountability as the binding constraints in practice. Vogel et al.'s finding deserves emphasis: in their study, the obstacles to effective AI adoption in analysis were organisational and epistemic — analysts could not interrogate the systems' reasoning and therefore could not calibrate trust in it — rather than technical. A substitution-sceptical school, anchored by Goldfarb and Lindsay (2022), reaches a convergent conclusion by economic reasoning: prediction machines make human judgement more valuable, not less, and the strategic risk lies in institutions mistaking cheap prediction for cheap judgement. A

transformation school (Horowitz, 2018; Payne, 2021) locates the significance of AI at the level of the international system and the character of conflict rather than the analytic workstation. And an organisational-limits school (Jensen et al., 2020) warns that all three preceding camps overestimate how readily bureaucracies metabolise the technology.

A fifth literature examines the threat side of the same systems. Goldstein et al. (2023) analyse how generative language models alter the economics of influence operations: content generation ceases to be a bottleneck, personalisation becomes cheap, and attribution becomes harder as machine-written text loses distinguishing marks. Buchanan (2020) frames the underlying dependency compactly — capability follows from data, algorithms and compute — which entails that whoever shapes the data shapes the capability, an entailment with obvious application to adversaries who can seed the open corpus. Nor is the seeding hypothetical: Greshake et al. (2023) demonstrated that adversarial instructions planted in retrieved content can hijack deployed language-model applications without any direct access to them — indirect prompt injection — and Carlini et al. (2024) showed that poisoning the web-scale corpora on which foundation models train is practical today at trivial cost, through expired-domain purchase and snapshot front-running. The model attack surface is demonstrated capability, not conjecture — and it has widened with the systems themselves. Retrieval-augmented generation, adopted precisely to ground model output in current sources, imports a corruption channel of its own: a handful of adversarial passages injected into a retrieval corpus reliably steers the answers generated from it (Zou et al., 2025). The agentic turn — models that plan, invoke tools and act across systems, increasingly through shared interconnection standards such as the Model Context Protocol (Anthropic, 2024) — multiplies the points at which planted instructions enter, and the benchmarks built to measure exactly this find current defences porous (Debenedetti et al., 2024; Zhan et al., 2024). Read together, these literatures expose an uncomfortable circularity that none of them theorises: intelligence agencies propose to analyse the open information environment using models trained on that environment, while adversaries manipulate the environment partly in order to manipulate the models. The augmentation school does not address model poisoning; the threat literature does not address analytic tradecraft; and no existing framework specifies where, in an intelligence architecture, the boundary between machine processing and human judgement should sit, or what properties the interface between them must have. These are the questions the framework developed in Section 4 is built to answer.

A further body of work, produced largely inside the model developers themselves, addresses the control of the systems rather than their use, and intelligence studies has yet to register it. Constitutional AI trains models against an explicit, published set of written principles rather than unstructured preference data alone (Bai et al., 2022), and behaviour-specification documents such as OpenAI's Model Spec make intended model conduct an inspectable artefact (OpenAI, 2024); red teaming has been industrialised from research method (Ganguli et al., 2022; Perez et al., 2022) into standing programmes whose accumulated operational lessons are now themselves published (Bullwinkel et al., 2025) and whose limits are under critical examination (Feffer et al., 2024); and dangerous-capability evaluation frameworks propose gating deployment on measured risk (Shevlane et al., 2023). None of this work concerns intelligence analysis, and all of it matters here: specifications, evaluations

and red teams are the raw material from which the model-governance and model-defence functions of Section 4.2.4 can actually be built, and theirs is the vocabulary in which analytic AI procurement and licensing will be written.

## 2.5 Decision analytics and evidence-based decision-making

Intelligence exists to serve decisions, and any theory of intelligence must therefore rest on a defensible account of how decisions are actually made. The relevant literature spans three research traditions whose relationship is uneasy. The first descends from Simon's (1955) demonstration that human rationality is bounded: decision-makers satisfice within the limits of attention, memory and computation rather than optimise over complete information. The second, the heuristics-and-biases programme, catalogued the systematic departures from normative judgement that bounded cognition produces — anchoring, availability, representativeness (Tversky & Kahneman, 1974) — and matured into a general account of fast and slow cognition (Kahneman, 2011). The third, naturalistic decision-making, studied experts in the field rather than students in the laboratory and found that experienced decision-makers under time pressure do not compare options at all; they recognise situations and act on the first workable course (Klein, 1998). The traditions clashed for decades before Kahneman and Klein (2009) mapped a partial reconciliation: intuitive expertise is trustworthy where the environment offers stable regularities and rapid feedback, and untrustworthy where it does not. Strategic intelligence, unhappily, is close to the paradigm of a low-validity environment — rare events, delayed and ambiguous feedback — which implies that analytic intuition alone cannot carry the enterprise.

A fourth strand tested judgement empirically at scale. Tetlock and Gardner (2015) showed that forecasting accuracy varies enormously between individuals, that it improves with training, teaming and calibration feedback, and that the best amateur forecasters outperformed professionals with access to classified material. Mandel and Barnes (2014), analysing a large corpus of real strategic intelligence forecasts, found analysts better calibrated than the pessimists expected but demonstrated that accuracy could be measured — and therefore managed — inside a working intelligence organisation. Dhami, Mandel, Mellers and Tetlock (2015) drew the programmatic conclusion: intelligence analysis should be rebuilt on decision science, with monitored accuracy, explicit probabilities and validated techniques, rather than on inherited craft lore. Kahneman, Sibony and Sunstein (2021) completed the error accounting: across judgement domains, noise — unwanted variability between equally qualified judges — typically contributes as much error as bias does, which strengthens the case for scored probability twice over, since scoring is the only discipline that measures both. The prescription remains largely unimplemented, and the literature is candid about the institutional reasons why.

The organisational wing of decision research asks where algorithmic systems fit. Duan, Edwards and Dwivedi (2019) survey the migration of AI from decision support to decision participation and identify the unresolved boundary question — which decisions may be delegated — as the field's central open problem. Shrestha, Ben-Menahem and von Krogh (2019) provide the most useful typology: full delegation to algorithms, hybrid sequential configurations in which one party filters options for the other, and aggregated human–machine judgement, each with distinct profiles of speed, interpretability and accountability. Jarrahi (2018) adds the epistemic rationale for hybridity: machines dominate where problems are computational, humans where they are equivocal — laden with conflicting interpretations rather than missing data. What the decision literature does not

supply is any engagement with adversarial information conditions. Its environments are noisy but honest; nobody in these models is manufacturing the evidence. For intelligence, that omission is disqualifying, and it is one the framework in Section 4 is designed to repair.

## 2.6 Knowledge management and information fusion

Two adjacent literatures address how organisations turn dispersed material into usable knowledge. Knowledge management theorised the organisational side. Nonaka and Takeuchi (1995) located competitive advantage in the conversion between tacit and explicit knowledge — socialisation, externalisation, combination, internalisation — a cycle with obvious resonance for analytic tradecraft, where the most valuable knowledge (source feel, target intuition, pattern memory) is overwhelmingly tacit. Alavi and Leidner (2001) systematised the field around knowledge processes — creation, storage, transfer, application — and issued a warning that intelligence organisations have repeatedly relearned: technology-centred knowledge systems fail where they ignore the social systems in which knowing occurs. Choo (1996) supplied the integrative account, the knowing organisation, in which sensemaking, knowledge creation and decision-making form a single continuous cycle. None of this work anticipated systems that generate knowledge claims themselves; its machines store and transmit, they do not assert. The question of how an organisation should absorb assertions produced by a statistical model — claims with no author, no tacit hinterland and no accountability — sits outside the theory's original scope.

Information fusion approached the same problem from engineering. Hall and Llinas (1997) codified multisensor data fusion, and the Joint Directors of Laboratories (JDL) model — revised by Steinberg, Bowman and White (1999) — remains the field's reference architecture, ascending from object refinement through situation and impact assessment to process refinement. The upper levels are, in all but name, intelligence analysis performed by machine. Khaleghi, Khamis, Karray and Razavi (2013), reviewing the state of the art, catalogue the unsolved difficulties with unusual candour: imperfect data, outliers and spurious inputs, conflicting sources — a problem class that becomes acute, and for which no principled method exists, where the conflict is engineered rather than accidental. The fusion literature thus offers something intelligence theory lacks, a formal layered architecture, while lacking something intelligence theory possesses, an appreciation that sources may be adversaries. The two literatures barely cite one another. Their integration — a fusion layer explicitly hardened against manipulation, feeding a human judgement layer — is one of the structural moves the present framework makes.

## 2.7 Sensemaking, situation awareness and cognitive bias

Where fusion describes machinery, sensemaking describes minds. Weick (1995) established the foundational inversion: organisations do not first perceive an environment and then interpret it; they enact interpretations retrospectively, driven more by plausibility than by accuracy. Klein, Moon and Hoffman (2006) sharpened the concept for macrocognition, defining sensemaking as the deliberate effort to fit data into a frame and the frame to data — a two-way contest in which the frame usually wins. Pirolli and Card (2005) gave the account its most operationally useful form for this field: their notional model of the analyst's

workflow as twin loops of foraging and sensemaking, each with characteristic costs and failure points, remains the best available description of what analysts actually do. Endsley's (1995) situation awareness model — perception, comprehension, projection — supplied the parallel account for time-critical domains and, crucially, theorised its own failure modes, including the automation-induced loss of awareness that arises when operators are placed out of the loop.

The bias literature explains why frames win. Heuer (1999) translated the heuristics-and-biases programme into intelligence doctrine: analysts perceive what they expect to perceive, mindsets form quickly and resist change, and new information is assimilated to existing beliefs. Nickerson (1998) documented the ubiquity of confirmation bias across domains and its resistance to warnings. Johnston (2005), studying the analytic community ethnographically from within, relocated the problem from the individual to the institution: time pressure, production quotas and a culture of precedent generate bias systematically, whatever the cognitive hygiene of individual analysts. The unresolved question this literature bequeaths to the AI era is uncomfortable and largely unexamined: machine systems trained on human-generated corpora do not stand outside these biases, they inherit them at scale — and add failure modes of their own, delivered with a fluency that human error never possessed. A sensemaking loop in which the foraging is machine-mediated and the frames are partly machine-suggested is a different cognitive object from the one Pirolli and Card described, and no adequate theory of it yet exists.

## 2.8 Structured analytic techniques

The intelligence community's institutional answer to cognitive bias has been the structured analytic technique. Heuer and Pherson (2021) codify the canon — analysis of competing hypotheses, key assumptions checks, devil's advocacy, red teaming and dozens more — on the rationale that externalising reasoning makes it inspectable, and inspectable reasoning is correctable. The rationale, however, ran ahead of the evidence. Coulthart (2017), evaluating the twelve core techniques against the empirical record, found validation studies scarce, methodologically weak and mixed in result; the community had mandated the techniques, particularly after the Iraq WMD post-mortems, largely on face plausibility. Chang, Berdini, Mandel and Tetlock (2018) pressed the critique to its sharpest point: some structured techniques may actively degrade judgement — decomposing a problem multiplies the surfaces on which bias can operate, and the ritual of structure can substitute for the substance of scepticism. Their proposed remedy is Bayesian in spirit: what needs structuring is not the pantomime of hypothesis tables but the disciplined updating of probability in the light of evidence. Marrin (2011) frames the whole dispute as symptomatic of a deeper professional failure — analysis has resisted becoming a discipline that tests its own methods.

For present purposes the debate matters twice over. First, it establishes that the human layer of any intelligence architecture cannot simply be asserted to provide 'judgement' as though judgement were a solved problem; the quality of human oversight is itself an empirical variable, and a fragile one. Second, and less noticed, the techniques were designed for a human cognitive budget. An analyst can check the key assumptions of a colleague's argument; no analyst can check the implicit assumptions of a model whose 'reasoning' is a

forward pass through billions of parameters. Whether structured techniques can be adapted to interrogate machine-generated analysis — a key assumptions check applied to a language model's summary, competing hypotheses run against retrieval pipelines — is a question the literature has not yet posed in earnest. The framework developed below treats it as central.

## 2.9 Explainable AI

If machine systems are to participate in analysis, their outputs must be assessable, and assessability is the province of explainable AI. The field's growth has been rapid and its internal disagreements instructive. Adadi and Berrada (2018) and Barredo Arrieta et al. (2020) map the territory: a proliferation of post-hoc techniques — feature attribution, surrogate models, counterfactual explanation — attached to models that remain opaque by construction. Miller (2019) delivered the field's most consequential critique from outside computer science: explanation is a social act with well-studied properties — contrastive, selective, conversational — and most XAI research ignores decades of relevant work in philosophy and psychology, producing explanations that satisfy developers rather than users. Rudin (2019) attacked from a different flank: for high-stakes decisions, post-hoc explanation of black-box models is the wrong project altogether, since such explanations are approximations that may misrepresent what the model actually computed; the correct demand is inherent interpretability. Doshi-Velez and Kim (2017) exposed a prior embarrassment — the field lacks agreed definitions or measures of interpretability itself — while Lipton (2018) showed the term conflates distinct desiderata (transparency, decomposability, simulatability, post-hoc justification) that different stakeholders want for different reasons. Gunning and Aha (2019), reporting DARPA's XAI programme, documented the practical corollary: explanation quality must be measured by its effect on user performance and trust calibration, not by algorithmic elegance.

The intelligence-specific stakes are higher than in most application domains, for three reasons the XAI literature registers only obliquely. First, analytic tradecraft standards require that judgements be traceable to sources and reasoning; a conclusion that cannot be interrogated cannot lawfully anchor a national assessment, whatever its accuracy. Second, the consumer of the explanation is not a lay user but a trained sceptic whose professional function is to probe reasoning — which makes Miller's (2019) social account of explanation directly applicable and the developer-centred alternatives directly inadequate. Third, under adversarial conditions explanations acquire a security function: an analyst who can see why a model flagged a pattern can also see when the model is being played. Vogel et al. (2021) found precisely this in the field — analysts declined to rely on systems whose reasoning they could not inspect, an instance of rational distrust rather than technophobia. The literature's unresolved question is whether explanation fidelity sufficient for tradecraft is achievable for

foundation models at all, or whether intelligence applications require Rudin's stronger medicine of interpretable-by-design architectures at the judgement-critical points of the pipeline.

## 2.10 Ethics of intelligence and of artificial intelligence

Two ethics literatures converge on the framework's subject matter from opposite directions. The ethics of intelligence, long an underdeveloped corner of the field, matured through Bellaby's (2014) harm-based account — intelligence activities are graduated intrusions requiring proportionate justification — and Omand and Phythian's (2018) sustained argument that just-war categories (right authority, proportionality, necessity, discrimination) can and should govern secret activity. Both accounts, however, were built for human collection against human targets; their categories strain when the 'activity' is continuous machine ingestion of a population's public expression. The ethics of AI, meanwhile, produced principles at industrial scale: Floridi et al. (2018) synthesised them into beneficence, non-maleficence, autonomy, justice and explicability, and Jobin, Ienca and Vayena (2019), reviewing eighty-four guideline documents, found global convergence on the vocabulary of principles accompanied by sharp divergence on their interpretation and near silence on their enforcement. Taddeo and Floridi (2018) named the resulting risk — ethics as decoration — and the intelligence application of these principles remains, as Blanchard and Taddeo (2023) document in a systematic review of the question, largely programmatic: the field knows that analytic AI raises problems of privacy aggregation, accountability for machine-informed judgement and proportionality at machine scale, and does not yet know what to do about them. Ghioni et al. (2024) reach the same verdict for OSINT specifically. The two literatures share a structural defect from this article's standpoint: both treat ethics as a constraint applied to intelligence from outside, rather than as a design property of the intelligence architecture itself — which is precisely how the governance layer of the present framework repositions it.

## 2.11 Human–machine teaming and hybrid intelligence

The oldest relevant finding in human factors research is also the most frequently rediscovered: automation changes the human role rather than removing it, and mismanaged automation degrades the very performance it was meant to improve. Parasuraman and Riley (1997) supplied the canonical taxonomy — use, misuse, disuse and abuse of automation — showing that operators both over-rely on automation that happens to be wrong and under-rely on automation that happens to be right. Lee and See (2004) reframed the problem as one of trust calibration: the design objective is neither trust nor suspicion but appropriate trust, matched to the system's actual reliability in the situation at hand. The behavioural literature then complicated the picture in both directions. Dietvorst, Simmons and Massey (2015) demonstrated algorithm aversion — people abandon algorithms after seeing them err, even when the algorithms remain superior to human judgement — while Logg, Minson and Moore (2019) demonstrated the opposite bias, algorithm appreciation, in which lay judges over-weight algorithmic advice. The pair of findings dissolves any simple story about human scepticism; which failure mode dominates is an empirical matter of task, expertise and presentation.

A younger literature attempts synthesis under the banner of hybrid intelligence. Dellermann, Ebel, Söllner and Leimeister (2019) define it as the combination of human and machine capabilities to achieve outcomes neither could reach alone, with complementarity — not substitution — as the design principle. Akata et al. (2020) set out the research agenda: collaboration, adaptivity, responsibility and explainability as joint requirements of systems built to work with people. Amershi et al. (2019) distilled the interaction findings into design guidelines whose empirical grounding exceeds most of the field. Rahwan et al. (2019) supplied the framing insight that machine behaviour must be studied as behaviour — empirically, in context, in interaction with humans and other machines — rather than deduced from code. Read against Sections 2.7 and 2.9, this literature furnishes the micro-foundations the framework requires: trust calibration presupposes inspectable reasoning (hence explainability is load-bearing, not cosmetic), and complementarity presupposes a principled division of epistemic labour (hence task allocation is a theoretical question, not an implementation detail). What the teaming literature lacks, once again, is an adversary: its systems fail statistically, not deliberately, and nobody in its experiments is manipulating the inputs to exploit the trust dynamics it so carefully measures. Intelligence work offers no such courtesy.

## 2.12 Intelligence governance, oversight and cognitive security

Governance scholarship in intelligence studies has concentrated on retrospective accountability: legislative committees, judicial warrants, inspectors-general. Born and Leigh (2005) codified the comparative legal framework, and Omand (2010) argued from practice that legitimacy is not a constraint on intelligence effectiveness but a condition of it — public licence, once forfeited, takes capabilities with it. The model underlying this literature assumes, however, that the consequential decisions occur at points a reviewer can later inspect: a warrant sought, an operation approved, a report issued. Machine-mediated analysis dissolves that assumption. When collection prioritisation, source weighting and first-draft assessment are performed continuously by models, the consequential 'decisions' are design decisions — training data selection, objective functions, retrieval configurations — made by engineers long before any overseeable analytic act occurs. Oversight that inspects finished assessments while ignoring the pipeline that produced them audits the shadow rather than the object. Vogel et al. (2021) identify exactly this displacement of accountability as a first-order problem for AI-enabled analysis; the oversight literature has yet to absorb it. The instruments that could absorb it already exist in adjacent law and standardisation: the European Union's Artificial Intelligence Act (Regulation (EU) 2024/1689) imposes design-time obligations — risk management, data governance, logging, human oversight — on high-risk systems; the NIST AI Risk Management Framework (National Institute of Standards and Technology, 2023) specifies governance functions for precisely the pipeline stages that retrospective review cannot reach; and ISO/IEC 42001:2023 renders AI management auditable as an organisational system (International Organization for Standardization, 2023). That intelligence studies has yet to engage this corpus — and that the corpus, in turn, largely exempts national-security systems from its scope — measures the same governance gap from both sides. Above statute and standard, a frontier-governance layer has formed with unusual speed: national AI safety institutes now run pre-deployment

evaluations of the most capable models, and the first International AI Safety Report consolidates the scientific evidence base behind that practice on behalf of thirty governments (Bengio et al., 2025); NIST has extended its framework with a generative-AI profile (National Institute of Standards and Technology, 2024); the EU has operationalised the AI Act's general-purpose obligations through a code of practice (European Commission, 2025); and proposals for regulating frontier development as such predate all of these (Anderljung et al., 2023). Implementation scholarship — how ISO/IEC 42001 audits perform in practice, whether safety-institute evaluations catch what matters — remains close to empty, which is itself a datum: the instruments are arriving faster than the evidence that they work. Beneath both sits a documentation stratum the framework can reuse directly: model cards standardise the reporting of a model's intended use and limits (Mitchell et al., 2019), datasheets do the same for training data (Gebru et al., 2021), system cards extend the practice to deployed products (OpenAI, 2023), incident cataloguing gives failures a shared registry (McGregor, 2021), and 'AI assurance' names the audit ecosystem forming around these artefacts (Department for Science, Innovation and Technology, 2024). These are the concrete forms the accountability architecture of Section 4.2.4 will take.

A separate literature grew up around the integrity of the information environment itself. Wardle and Derakhshan (2017) provided the standard taxonomy of information disorder — misinformation, disinformation, malinformation — and the insistence that intent and falsity are independent dimensions. Benkler, Faris and Roberts (2018) demonstrated empirically that disorder propagates through structural features of media ecosystems rather than through technology alone, and Rid (2020) supplied the historical corrective: active measures are a century-old professional practice whose digital forms are continuous with, not a rupture from, their analogue ancestry. Bradshaw and Howard (2019) documented the industrialisation — organised social media manipulation as standing state capability across dozens of countries. Peer-reviewed synthesis converged on the same behavioural turn: Freelon and Wells (2020) frame disinformation as the defining problem of contemporary political communication, and Starbird, Arif and Wilson (2019) show that information operations function as participatory, collaborative work within online crowds — which is exactly why content-level adjudication cannot identify them. On the defensive side of the same contest, media-forensic detection of synthetic content is a mature field candid about its own limits (Verdoliva, 2020); provenance standards such as Content Credentials attack the problem from the authenticity side (Coalition for Content Provenance and Authenticity, 2023); and psychological inoculation — pre-exposure to weakened manipulation techniques — has shown measurable resilience effects at platform scale (Roozenbeek et al., 2022). These are input-defence and analyst-defence instruments before the framework names them as such. On the defensive side, Pamment (2020) and the institutional work culminating in the EEAS's (2023) first threat report on foreign information manipulation and interference shifted the analytic unit from content to behaviour: FIMI is identified by patterns of coordinated manipulative conduct, catalogued in shared taxonomies of adversary techniques, rather than by adjudicating truth. NATO's parallel concept development frames the contest at the cognitive level directly (du Cluzel, 2021). This literature has produced, in effect, an intelligence discipline of its own — with collection, analysis, taxonomies and countermeasures — while remaining institutionally and theoretically detached from

intelligence studies. The detachment is the point at which the present framework intervenes: if the open information environment is simultaneously the principal collection substrate (Section 2.3), the training corpus of the analytic machinery (Section 2.4) and the adversary's preferred attack surface, then cognitive security is not adjacent to intelligence production. It is a layer of it.

# 3. The Theoretical Gap

The review supports a diagnosis more specific than the familiar complaint that intelligence studies lacks theory. The field possesses theory in abundance — process models, epistemological critiques, decision science, fusion architectures, teaming principles, governance frameworks. What it lacks is integration at the point where the contemporary problem actually lives: the joint operation of machine analysis, open-source dominance and adversarial information conditions. Five gaps, each visible only when the literatures are read against one another, define the space the framework must fill.

The first is an integration gap. The literatures reviewed above are severally sophisticated and mutually isolated. Fusion engineering has a layered architecture but no adversary theory; the disinformation field has an adversary theory but no seat inside intelligence process models; XAI has explanation techniques but no connection to analytic tradecraft standards; the teaming literature has trust dynamics but assumes honest inputs; intelligence theory has process critique but no machine epistemology. No existing framework provides the connective tissue, and the problem is not divisible: an architecture that solves trust calibration while ignoring input poisoning solves nothing, because poisoned inputs are exactly how an adversary exploits calibrated trust.

The second is a division-of-labour gap. Every reviewed literature that touches human–machine configuration agrees that the allocation of analytic tasks between humans and machines matters decisively — and none specifies the allocation for intelligence work. Shrestha et al. (2019) offer configurations without domain content; Jarrahi (2018) offers a criterion (computational versus equivocal problems) without an intelligence mapping; Goldfarb and Lindsay (2022) establish that judgement gains value without locating where judgement must sit in a production pipeline. The question 'where, precisely, does machine processing end and human judgement begin, and what must the interface guarantee?' has no answer in the current literature. It is the framework's second research question and one of its principal deliverables.

The third is an adversarial-input gap, and it is the most dangerous of the five. The decision, fusion, teaming and XAI literatures uniformly assume environments that are noisy but not hostile: errors are random, sources are honest, and the evidence base is not itself a battlefield. Intelligence enjoys no such assumption, and the modern configuration compounds the exposure into a circularity no literature has theorised (Figure 1): agencies analyse the open environment with models trained on that environment, while adversaries manipulate the environment partly to manipulate the models. Khaleghi et al. (2013) list conflicting and spurious inputs among fusion's unsolved problems — a class that adversarial deception industrialises; Goldstein et al. (2023) show the industrialisation under way; and the two observations have not, to the author's knowledge, been placed inside a single theory of intelligence production.

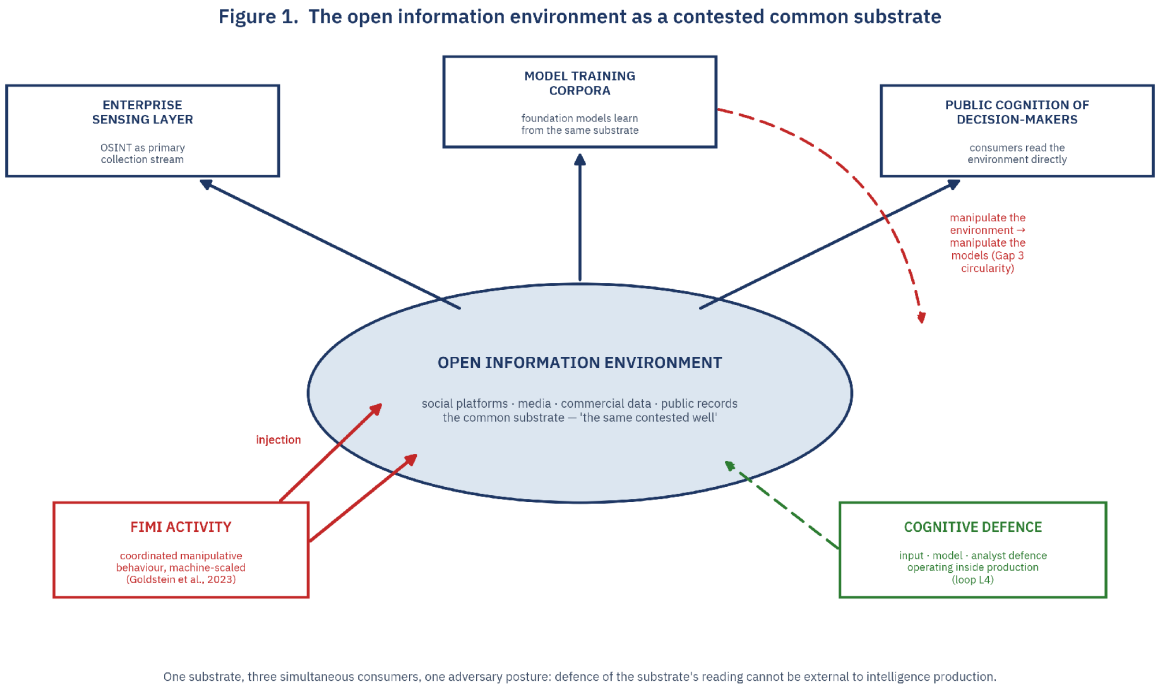


Figure 1. The open information environment as a contested common substrate
ENTERPRISE SENSING LAYER
OSINT as primary collection stream
MODEL TRAINING CORPORA
foundation models learn from the same substrate
PUBLIC COGNITION OF DECISION-MAKERS
consumers read the environment directly
manipulate the environment → manipulate the models (Gap 3 circularity)
OPEN INFORMATION ENVIRONMENT
social platforms · media · commercial data · public records
the common substrate — 'the same contested well'
injection
FIMI ACTIVITY
coordinated manipulative behaviour, machine-scaled (Goldstein et al., 2023)
COGNITIVE DEFENCE
input · model · analyst defence operating inside production (loop L4)
One substrate, three simultaneous consumers, one adversary posture: defence of the substrate's reading cannot be external to intelligence production.

**Figure 1. The open information environment as a contested common substrate**

The fourth is an explainability–tradecraft gap. Intelligence doctrine requires that judgements be traceable to sources and reasoning; XAI cannot yet guarantee that its explanations correspond to what models actually compute (Rudin, 2019), and cannot even define the property being demanded (Lipton, 2018). The consequence is that machine-drafted analysis currently enters intelligence production, where it enters at all, on terms no human product would be permitted: fluent, confident and unaccountable. The reviewed literatures name the two halves of the problem; none joins them into a requirement specification for analytic AI.

The fifth is a governance-timing gap. Oversight theory audits acts; machine-mediated analysis relocates the consequential choices to design time, where existing oversight does not reach (Vogel et al., 2021). Both ethics literatures treat their subject as an external constraint on a finished system rather than a property engineered into the system. A theory adequate to the present condition must specify governance as architecture — provenance, auditability, documentation, red-teaming — embedded where the decisions actually occur.

If we invert the problem, what the discipline actually needs today is not another isolated taxonomy, but a unifying architecture. We require a framework that refuses to treat machine sensing, human judgement, and governance as separate phases, forcing them instead into a single, cohesive structure in which (i) machine sensing, information fusion, human judgement and governance are distinct, specified layers; (ii) the human–machine boundary is theorised by comparative epistemic advantage; (iii) explainability is a load-bearing interface requirement rather than a feature; (iv) cognitive defence against manipulation of inputs, models and analysts is structural; and (v) governance operates at design time. Intelligence Theory 2026, developed in the following section, is constructed to those five

specifications. Its originality claim rests not on any single component — each has a parent literature — but on the demonstration that the components are jointly necessary and on the specification of their relations, feedback loops and testable consequences.

**Table 9. Gap matrix: what each literature supplies and what it omits**

| Literature | Layered architecture | Human–machine boundary | Adversarial inputs | Explainability–tradecraft link | Design-time governance |
|---|---|---|---|---|---|
| Intelligence process theory (§2.1) | Partial (cycle stages) | Absent | Partial (deception lore) | Absent | Absent |
| Strategic / military AI (§2.2, §2.4) | Absent | Partial (complementarity) | Partial (threat-side only) | Absent | Absent |
| OSINT (§2.3) | Absent | Absent | Partial (verification problem named) | Absent | Partial (ethics critique) |
| Decision analytics (§2.5) | Absent | Partial (configurations) | Absent | Absent | Absent |
| Information fusion (§2.6) | Present (JDL) | Absent | Named as unsolved | Absent | Absent |
| Sensemaking / bias / SATs (§2.7–2.8) | Absent | Absent | Absent | Partial (reasoning externalised) | Absent |
| Explainable AI (§2.9) | Absent | Absent | Absent | Partial (techniques, no doctrine) | Absent |
| Ethics (§2.10) | Absent | Absent | Absent | Partial (explicability principle) | Partial (principles, no architecture) |
| Human–machine teaming (§2.11) | Absent | Present (complementarity) | Absent | Partial (trust calibration) | Absent |
| Governance / cognitive security (§2.12) | Absent | Absent | Present (FIMI/DISARM) | Absent | Partial (retrospective oversight) |

The matrix makes the diagnosis legible at a glance: no column is empty, and no row fills more than two columns. The materials for an adequate theory exist, distributed across ten literatures; the theory does not. Section 4 assembles it.

# 4. The Conceptual Framework: Intelligence Theory 2026

## 4.1 Orientation and core construct

The framework developed in this section answers the five specifications derived at the close of Section 3. Its architecture can be stated in one sentence before being unpacked at length: intelligence production is modelled as four functionally distinct, mutually regulating layers — sensing, fusion, judgement and governance — in which analytic labour is allocated between humans and machines by comparative epistemic advantage, every machine-to-human interface carries an explainability obligation, and the protection of the enterprise's own inputs, models and analysts from adversarial manipulation is a permanent function of the architecture rather than an occasional counter-intelligence task. The framework as a whole is termed Intelligence Theory 2026; its organising construct is Defensive Hybrid Intelligence.

**Definition 1 (Defensive Hybrid Intelligence).** Defensive Hybrid Intelligence (DHI) is the organised production of assessed knowledge for decision through the structured collaboration of human and machine analytic capabilities, under an explicit and continuously exercised regime of epistemic self-protection against adversarial manipulation of the enterprise's information inputs, computational models and human cognition.

This formulation is highly deliberate. The components of this definition are not interchangeable academic buzzwords; they represent strict operational constraints that the architecture must satisfy 'Organised production of assessed knowledge for decision' retains what is durable in the classical tradition from Kent (1949) through Herman (1996): intelligence is judgement in the service of action, not information for its own sake. 'Structured collaboration of human and machine analytic capabilities' registers the hybrid condition established in Sections 2.4 and 2.11 — machines are participants in analysis, not tools beneath it — while 'structured' signals that the collaboration is a designed allocation, not an ad hoc accretion. 'Explicit and continuously exercised regime of epistemic self-protection' is the defensive element, and its three named targets — inputs, models, cognition — correspond to the three attack surfaces (Figure 2) that Sections 2.3, 2.4 and 2.12 showed to be jointly exposed: the open-source substrate can be poisoned, the models trained on it can be manipulated through it, and the analysts reading both can be targeted directly. A hybrid intelligence without the defensive regime reproduces the vulnerability circularity identified as Gap 3; a defensive posture without hybrid structure forfeits machine scale. The construct requires both.

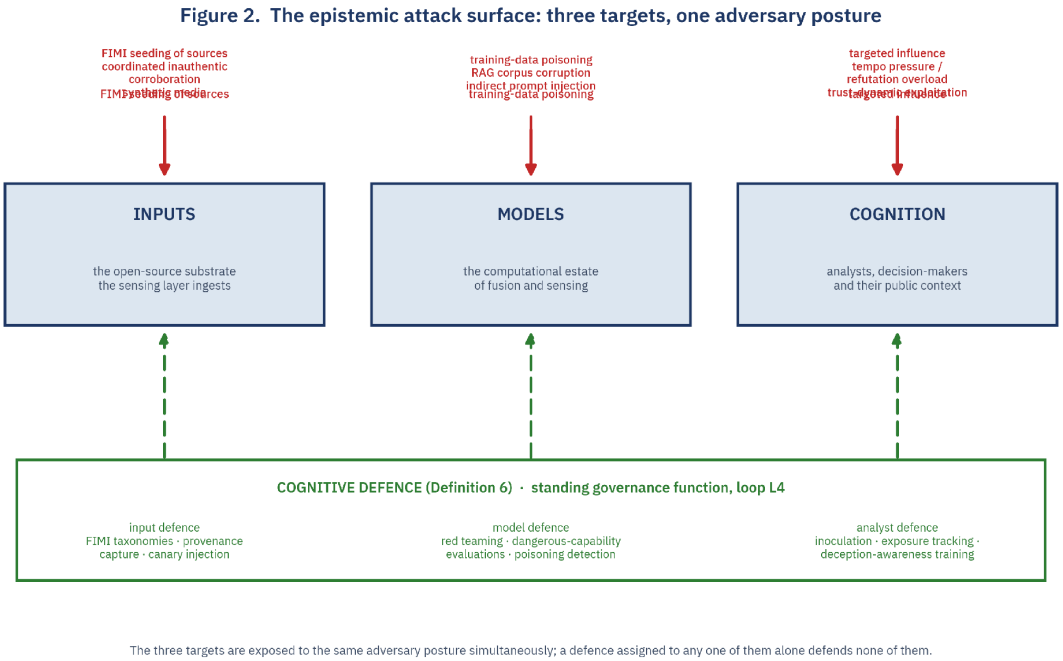


Figure 2. The epistemic attack surface: three targets, one adversary posture
FIMI seeding of sources
coordinated inauthentic corroboration
training-data poisoning
RAG corpus corruption
indirect prompt injection
targeted influence
tempo pressure / refutation overload
INPUTS
the open-source substrate the sensing layer ingests
MODELS
the computational estate of fusion and sensing
COGNITION
analysts, decision-makers and their public context
COGNITIVE DEFENCE (Definition 6) · standing governance function, loop L4
input defence
FIMI taxonomies · provenance capture · canary injection
model defence
red teaming · dangerous-capability evaluations · poisoning detection
analyst defence
inoculation · exposure tracking · deception-awareness training
The three targets are exposed to the same adversary posture simultaneously; a defence assigned to any one of them alone defends none of them.

**Figure 2. The epistemic attack surface: three targets, one adversary posture**

Two boundary clarifications prevent misreading. First, DHI is not a technology stack. It is a theory of organisation: it specifies functions, allocations, interfaces and obligations, and is agnostic about which particular systems discharge them. Second, 'defensive' does not mean passive. The regime of epistemic self-protection includes active measures — adversarial testing of models, deliberate injection of known-false canaries into pipelines to measure contamination, red-teaming of analytic lines — on the counter-intelligence principle that a defence never exercised is a defence never verified.

## 4.2 The four layers

The layers are functional, not organisational (Figure 3): they describe what must be done and by what class of agent, not which office does it. A small cell and a national agency instantiate the same four functions at different scales. The layers are presented bottom-up, following the direction of primary information flow, though — as Section 4.5 establishes — the feedback flows running downward are what distinguish a living architecture from a pipeline.

Figure 3. Intelligence Theory 2026: the Defensive Hybrid Intelligence architecture

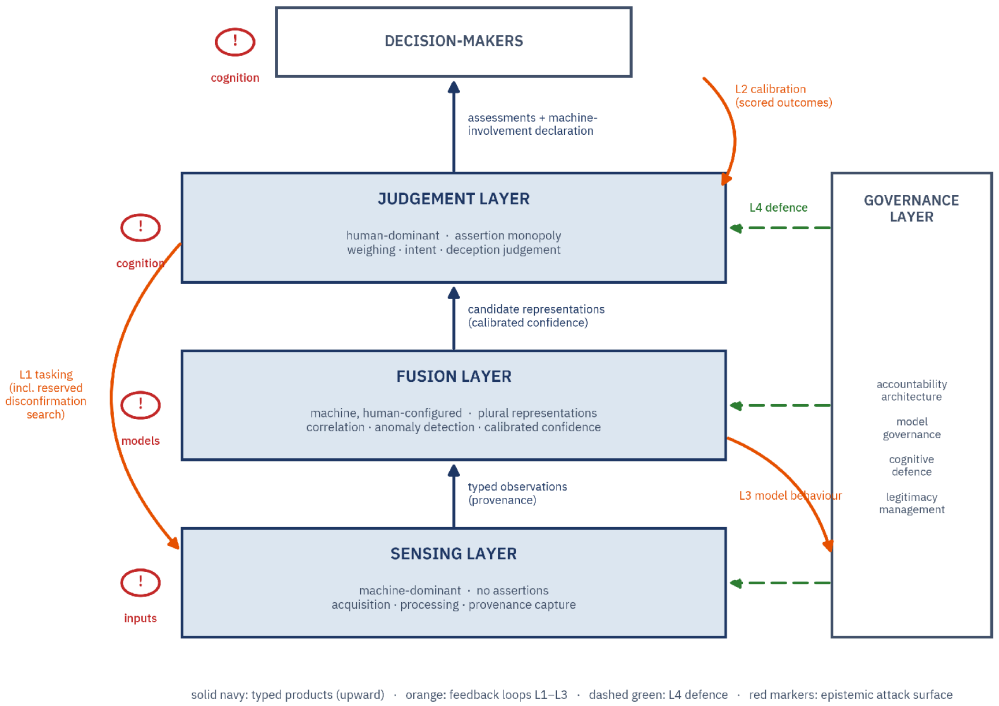



**Figure 3. Intelligence Theory 2026: the Defensive Hybrid Intelligence architecture**

### *4.2.1 The sensing layer*

The sensing layer performs continuous, machine-dominant acquisition and first-order processing of the information environment: collection across open and controlled sources, ingestion, translation, entity extraction, deduplication, format normalisation and first-pass triage. Its defining characteristics are scale and statelessness — it operates at volumes no human staff can address (Zegart, 2022), and it asserts nothing. That second property is a design rule, not an observation: outputs of the sensing layer are typed as observations with provenance, never as claims. The distinction matters because the layer is machine-dominant and machines inherit corpus bias (Bommasani et al., 2021); permitting sensing outputs to enter the enterprise as assertions would launder statistical artefact into evidence. Provenance capture is therefore obligatory at this layer and immutable thereafter: every item carries its origin, acquisition time, transformation history and collection rationale forward through the architecture. Within the classical vocabulary, the sensing layer absorbs collection and processing; within the fusion vocabulary, JDL levels 0 and 1 (Steinberg et al., 1999). Unlike either ancestor, it treats the open information environment as its primary substrate and therefore as a contested space, which is why it operates under the permanent supervision of the governance layer's input-defence function (Section 4.2.4).

### *4.2.2 The fusion layer*

The fusion layer converts typed observations into candidate situational representations: correlation across sources and modalities, network reconstruction, geospatial and temporal

alignment, anomaly detection, and machine-drafted situation summaries. It corresponds to JDL levels 2 and 3 (Steinberg et al., 1999) and is the layer where large language models and allied systems do their most consequential work — and where, accordingly, the framework imposes its most consequential constraints. Three are definitional. First, plurality: the fusion layer must maintain competing representations rather than converging on one, because premature fusion is the computational form of premature closure (Heuer, 1999), and because a single fused picture presents a single target to an adversary shaping inputs. Second, calibrated uncertainty: every fusion product carries explicit confidence representation, propagated from source reliability scores rather than asserted post hoc. Third, contestability: every fusion product must be decomposable on demand into the observations and transformations that produced it — the architectural precondition for the explainability obligation defined in Section 4.4. A fusion product that cannot be decomposed is, by definition within this framework, not admissible to the judgement layer. This is the point at which Rudin's (2019) argument acquires architectural force: at judgement-critical junctions, interpretable-by-design methods are preferred to post-hoc explanation of opaque ones, accepting accuracy costs where necessary, because the layer's outputs must survive interrogation by trained sceptics, not merely satisfy developers (Miller, 2019).

### *4.2.3 The judgement layer*

The judgement layer is human-dominant and produces the enterprise's assertions: assessments, estimates, warnings, and the confidence attached to them. Its function is not to repeat the fusion layer's work at lower speed but to perform the operations for which humans hold comparative epistemic advantage (Jarrahi, 2018; Goldfarb & Lindsay, 2022): weighing equivocal evidence, reasoning about adversary intent through theory of mind, judging deception, taking responsibility, and deciding what the assessment is for — the Kendall (1949) question of relevance to the decision-maker's actual problem. The layer's tradecraft is structured but reformed structure: following Chang et al. (2018) and Dhami et al. (2015), its core disciplines are explicit probability, recorded rationale, tracked accuracy and systematic exposure to disconfirmation, rather than technique ritual. Two design rules govern its relationship to the machine layers below. The first is the assertion monopoly: only the judgement layer asserts. Machine-drafted text may inform an assessment; it may not be one, however fluent, because assertion within this framework entails accountability and machines cannot bear it (Vogel et al., 2021). The second is mandatory interrogation: no fusion product is accepted into an assessment without documented challenge — a key-assumptions check run against the machine's representation, an explicit consideration of how the product would look if its principal sources were adversarially seeded. This is the adaptation of structured techniques to machine outputs whose absence Section 2.8 noted; the framework makes it a standing obligation rather than an analyst's option, precisely because the trust-calibration literature (Lee & See, 2004; Logg et al., 2019) predicts that fluent machine products will otherwise acquire unearned authority.

### *4.2.4 The governance layer*

The governance layer is not above the other three in a supervisory diagram drawn after the fact; it is woven through them at design time, which is the repair of Gap 5. It discharges four standing functions. Accountability architecture maintains the provenance, documentation

and audit machinery that makes every assessment reconstructable: which observations, which models under which versions, which human judgements, in which order. Model governance controls the machine estate — validation before deployment, behavioural monitoring in deployment (Rahwan et al., 2019), version discipline, and documented boundaries of use for each system. Cognitive defence operates the DHI regime proper across the three attack surfaces (Figure 4): input defence (contamination detection and source-manipulation analysis in the sensing layer, drawing directly on FIMI behavioural taxonomies; EEAS, 2023; Pamment, 2020), model defence (adversarial testing, poisoning detection, canary injection), and analyst defence (exposure tracking, deception-awareness training, protection of the judgement layer from targeted influence). Legitimacy management maintains the ethical and legal envelope — proportionality of collection, privacy aggregation limits, oversight interfaces — as an engineered property rather than an external review, operationalising Omand's (2010) argument that public licence is a condition of capability and answering the critique of externally applied ethics developed in Section 2.10.

Figure 4. Cognitive defence: instruments and propagation paths of loop L4

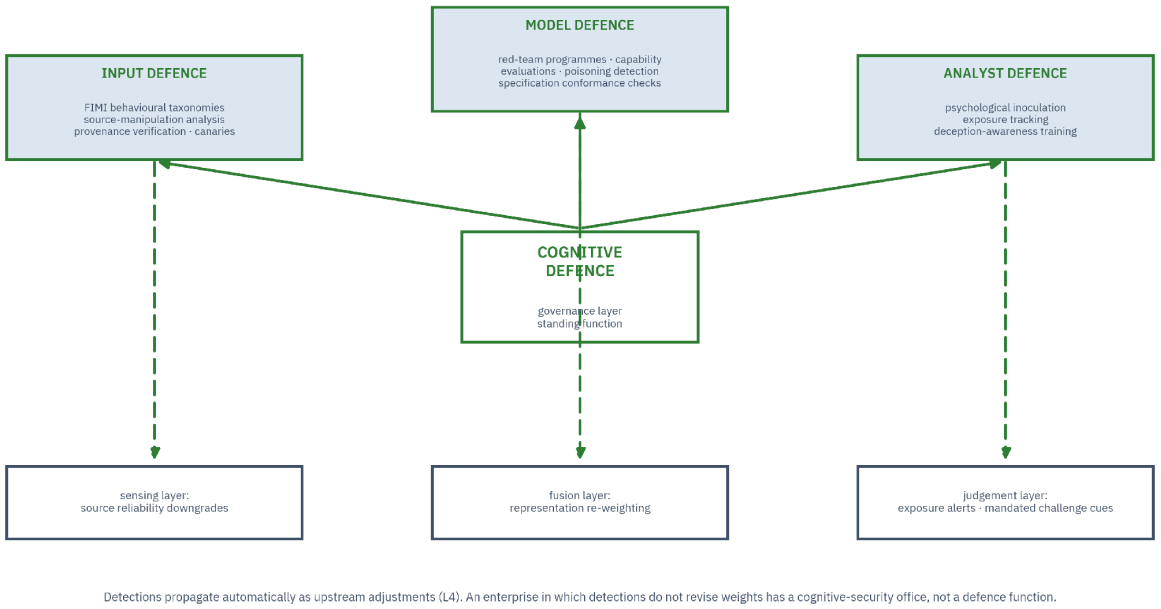

MODEL DEFENCE
red-team programmes · capability evaluations · poisoning detection specification conformance checks
INPUT DEFENCE
FIMI behavioural taxonomies source-manipulation analysis provenance verification · canaries
ANALYST DEFENCE
psychological inoculation exposure tracking deception-awareness training
COGNITIVE DEFENCE
governance layer standing function
sensing layer: source reliability downgrades
fusion layer: representation re-weighting
judgement layer: exposure alerts · mandated challenge cues
Detections propagate automatically as upstream adjustments (L4). An enterprise in which detections do not revise weights has a cognitive-security office, not a defence function.

**Figure 4. Cognitive defence: instruments and propagation paths of loop L4**

**Table 10. The four layers: functions, dominant agents and characteristic failure modes**

| Layer | Core functions | Dominant agent | Output type | Characteristic failure modes |
|---|---|---|---|---|
| Sensing | Acquisition, processing, translation, extraction, triage, provenance capture | Machine | Typed observations with provenance (no assertions) | Coverage gaps; ingestion of poisoned sources; provenance loss |

| Layer | Core functions | Dominant agent | Output type | Characteristic failure modes |
|---|---|---|---|---|
| Fusion | Correlation, network and situation reconstruction, anomaly detection, drafting | Machine, human-configured | Competing candidate representations with calibrated confidence | Premature convergence; opaque products; confidence miscalibration |
| Judgement | Weighing, intent assessment, deception judgement, assertion, relevance decisions | Human, machine-supported | Assessments and estimates with recorded rationale and probability | Automation bias; algorithm aversion; unchallenged machine drafts |
| Governance | Accountability architecture, model governance, cognitive defence, legitimacy management | Human, machine-instrumented | Standards, audits, defence operations, licences of use | Post-hoc drift; oversight of products but not pipelines; decorative ethics |

## 4.3 Construct definitions

Beyond DHI itself (Definition 1) and the four layers, the framework rests on six further constructs. They are defined formally here (Figure 5; Table 11) and used without further gloss in the remainder of the article.

**Definition 2 (Comparative epistemic advantage).** The principle that analytic tasks are allocated to the agent class — human or machine — whose characteristic capabilities and characteristic failure modes are respectively best and least suited to the task's epistemic demands, assessed under adversarial conditions rather than benign ones.

**Definition 3 (Assertion monopoly).** The design rule reserving the act of assertion — the issuing of an assessment for which the enterprise is accountable — to the human judgement layer. Machine outputs are observations or representations; they become assertions only through accountable human adoption.

**Definition 4 (Explainability obligation).** The requirement that every machine product crossing into the judgement layer be decomposable, on demand and within operational time, into its contributing observations, transformations and confidence derivations, to a standard sufficient for a trained analyst to locate and test its weakest dependency.

**Definition 5 (Epistemic attack surface).** The set of points at which an adversary can degrade the enterprise's knowledge production without penetrating it: the open-source inputs it senses, the models it runs, and the cognition of its analysts and consumers.

**Definition 6 (Cognitive defence).** The standing governance function that monitors, tests and hardens the epistemic attack surface, using behavioural taxonomies of manipulation for input defence, adversarial evaluation for model defence, and exposure management and training for analyst defence.

**Definition 7 (Trust calibration interface).** The designed presentation of machine products to human analysts such that displayed confidence, provenance and known failure

modes support reliance proportional to actual reliability (Lee & See, 2004), counteracting both automation bias and algorithm aversion.

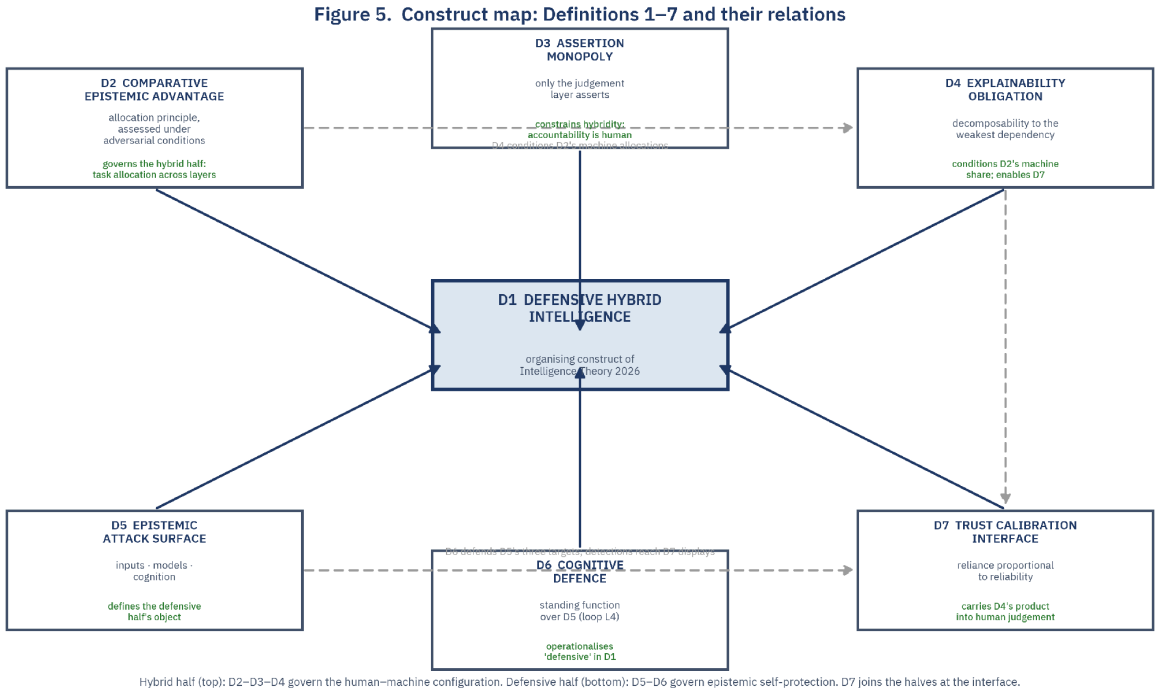


Figure 5. Construct map: Definitions 1–7 and their relations
D3 ASSERTION MONOPOLY
only the judgement layer asserts
constrains hybridity: accountability is human
D2 COMPARATIVE EPISTEMIC ADVANTAGE
allocation principle, assessed under adversarial conditions
governs the hybrid half: task allocation across layers
D4 EXPLAINABILITY OBLIGATION
decomposability to the weakest dependency
conditions D2's machine share; enables D7
D1 DEFENSIVE HYBRID INTELLIGENCE
organising construct of Intelligence Theory 2026
D5 EPISTEMIC ATTACK SURFACE
inputs · models · cognition
defines the defensive half's object
D6 COGNITIVE DEFENCE
standing function over D5 (loop L4)
operationalises 'defensive' in D1
D7 TRUST CALIBRATION INTERFACE
reliance proportional to reliability
carries D4's product into human judgement
Hybrid half (top): D2–D3–D4 govern the human–machine configuration. Defensive half (bottom): D5–D6 govern epistemic self-protection. D7 joins the halves at the interface.

**Figure 5. Construct map: Definitions 1–7 and their relations**

**Table 11. Construct summary: definitions, contrast cases and parent literatures**

| Construct | Definitional kernel | Contrast case (what it is not) | Parent literatures |
|---|---|---|---|
| D1 Defensive Hybrid Intelligence | Accountable human–machine knowledge production under a standing regime of epistemic self-protection | A technology stack; a passive security posture | Intelligence theory (§2.1–2.2); hybrid intelligence (§2.11); cognitive security (§2.12) |
| D2 Comparative epistemic advantage | Allocation by best capability and least exploitable failure mode, assessed under adversarial conditions | Benign-environment complementarity; allocation by procurement momentum | Jarrahi (2018); Goldfarb & Lindsay (2022); Dellermann et al. (2019) |
| D3 Assertion monopoly | Only the judgement layer issues assessments for which the enterprise is accountable | Machine text adopted by fluency; diffuse joint outputs | Kent–Kendall accountability tradition; Vogel et al. (2021) |
| D4 Explainability obligation | Decomposability of machine products to the weakest dependency within the decision class's window | Global model transparency; developer-satisfying post-hoc display | Rudin (2019); Miller (2019); Lipton (2018) |
| D5 Epistemic attack surface | Inputs, models and cognition as jointly exposed degradation points | Perimeter or network attack surface; content-level falsity | Goldstein et al. (2023); Carlini et al. (2024); EEAS (2023) |
| D6 Cognitive defence | Standing governance function monitoring, testing and hardening D5 (loop L4) | An adjacent counter-disinformation office; episodic audits | FIMI/DISARM tradition; red-teaming and evaluation literature (§2.4) |

| Construct | Definitional kernel | Contrast case (what it is not) | Parent literatures |
|---|---|---|---|
| D7 Trust calibration interface | Presentation supporting reliance proportional to actual reliability | Confidence display as decoration; persuasion toward acceptance | Lee & See (2004); Dietvorst et al. (2015); Logg et al. (2019) |

**Table 12. Human–machine task allocation by comparative epistemic advantage**

| Analytic task | Allocation | Rationale | Residual obligation on the other agent |
|---|---|---|---|
| Volume triage, translation, entity extraction | Machine | Scale and speed; low equivocality (Jarrahi, 2018) | Human spot-audit against provenance; sampling for drift |
| Pattern and anomaly detection across sources | Machine | Statistical sensitivity beyond human attention | Human interrogation of flagged patterns for adversarial seeding |
| Situation drafting and summarisation | Machine | Throughput; first-draft economy | Mandatory challenge before adoption; assertion monopoly |
| Deception judgement and intent assessment | Human | Theory of mind; adversarial reasoning (Payne, 2021) | Machine supplies base rates, timelines and contradiction maps |
| Weighing equivocal or conflicting evidence | Human | Equivocality is interpretive, not computational | Machine maintains competing representations and confidence propagation |
| Assertion, estimation and warning | Human | Accountability cannot be delegated (Vogel et al., 2021) | Machine preserves full decision provenance for audit |
| Relevance to the decision problem | Human | The Kendall question is political, not technical | Machine tracks consumer feedback signals for the tasking loop |

## 4.4 Interfaces: explainability as a load-bearing requirement

Classical process models treat the joints between stages as conveyances: products move from collection to analysis to dissemination, and the theory has nothing to say about the moving. In the present framework the interfaces are where the theory concentrates, because the layer boundaries are also the boundaries between agent classes, and Section 2 established that this is precisely where the known pathologies live — automation misuse at the machine-to-human joint (Parasuraman & Riley, 1997), relevance failure at the human-to-consumer joint (Kendall, 1949; Dahl, 2013), tasking failure at the consumer-to-sensing joint (Hulnick, 2006).

The machine-to-human interface carries the explainability obligation (Definition 4), and its formulation is deliberately more demanding than most of the XAI literature in one respect and less demanding in another. More demanding: decomposability must reach the weakest dependency, because the analyst's professional function is to find where an argument breaks, and under adversarial conditions the weakest dependency is where manipulation enters — the explanation is a security instrument, not a courtesy (Section 2.9). Less demanding: the obligation does not require global transparency of model internals, which for foundation

models is unavailable (Lipton, 2018); it requires product-level accountability — this summary rests on these sources, weighted thus, transformed so — which is achievable through architecture (retrieval traceability, confidence propagation, interpretable junction models per Rudin, 2019) even where the underlying generator is opaque. The framework thereby converts an unresolved philosophical dispute about explanation into a specifiable engineering requirement with a tradecraft rationale, which is what Gap 4 demanded.

The same interface carries the trust calibration function (Definition 7). The behavioural findings of Section 2.11 imply that presentation is not neutral: identical products differently framed produce over-reliance or abandonment (Dietvorst et al., 2015; Logg et al., 2019). The interface therefore displays reliability information as first-class content — historical accuracy of the producing pipeline on comparable tasks, known failure modes, contamination indicators from cognitive defence — so that the analyst's reliance has something to calibrate against. The human-to-consumer interface, by contrast, carries the classical obligations in modernised form: assessments travel with explicit probability, recorded rationale and — a DHI addition — a machine-involvement declaration stating which layers and systems contributed, so that the consumer's trust can be calibrated as well as the analyst's.

## 4.5 Feedback loops

Four feedback loops convert the layered structure from a pipeline into an adaptive system, and Proposition 6 asserts that their quality dominates component sophistication in determining enterprise performance. They are specified here with their direction, signal and characteristic corruption, since — as Table 5 noted at JDL level 4 — feedback paths are themselves attack surfaces.

**L1, the tasking loop (judgement → sensing).** Analytic gaps and falsification needs recalibrate collection priorities continuously, replacing the fiction of consumer-originated requirements (Hulnick, 2006) with an explicit, logged retasking signal. Its characteristic corruption is confirmation-seeking: a judgement layer that tasks collection to support its existing line converts the loop into an amplifier of premature closure. The framework's counter is structural — a fixed fraction of retasking capacity is reserved for disconfirmation search, and the allocation is auditable.

**L2, the calibration loop (outcomes → judgement).** Assessed probabilities are scored against outcomes wherever outcomes are observable (Mandel & Barnes, 2014; Tetlock & Gardner, 2015), and the scores feed analyst training, technique selection and confidence discipline. Its corruption is gaming — forecasting only the scoreable — and its counter is governance ownership of the scoring rules.

**L3, the model-behaviour loop (deployment → model governance).** Deployed systems are monitored as behaving agents (Rahwan et al., 2019): drift, error clustering, and divergence between explained and actual behaviour trigger revalidation or withdrawal. Its corruption is silent degradation under distribution shift — the failure mode of every unmonitored model estate.

**L4, the defence loop (cognitive defence → all layers).** Detected manipulation attempts propagate as adjustments everywhere: source reliability downgrades in sensing, representation re-weighting in fusion, exposure alerts in judgement, and taxonomy updates in defence itself (EEAS, 2023). L4 is the loop that makes the architecture defensive in more than name; an enterprise in which manipulation detections do not automatically revise upstream weights has a cognitive-security office, not a cognitive-defence function.

**Table 13. Feedback loops: signal, effect and characteristic corruption**

| Loop | Direction | Signal | Adaptive effect | Characteristic corruption |
|---|---|---|---|---|
| L1 Tasking | Judgement → Sensing | Analytic gaps; falsification needs | Collection follows the state of the argument | Confirmation-seeking retasking |
| L2 Calibration | Outcomes → Judgement | Scored forecast accuracy | Trained judgement; disciplined confidence | Gaming towards scoreable questions |
| L3 Model behaviour | Deployment → Governance | Drift, error clusters, explanation divergence | Revalidation or withdrawal of systems | Silent degradation; alert fatigue |
| L4 Defence | Cognitive defence → All layers | Detected manipulation patterns | Automatic upstream re-weighting; hardened surface | Loop itself spoofed to induce false downgrades |

## 4.6 Decision architecture

The framework's decision architecture specifies how the layered production connects to the strategic decisions it serves, and it does so by typing decisions rather than by prescribing a single flow (Figure 6). Following the configurational logic of Shrestha et al. (2019) and the environmental logic of Kahneman and Klein (2009), decisions supported by the enterprise are classed on two dimensions: tempo (deliberative to time-critical) and equivocality (computational to interpretive). The class determines the legitimate configuration. Deliberative-interpretive decisions — strategic estimates, warning judgements — take the full path: plural fusion representations, mandated challenge, human assertion, full provenance. Time-critical-computational decisions — indications monitoring, technical alerting — may run machine-to-consumer on pre-authorised lanes, but only for decision types explicitly licensed by governance in advance, with the licence itself documented, bounded and periodically re-validated: automation of the act, never of the authority. Time-critical-interpretive decisions — the hardest class — receive the framework's most distinctive treatment: pre-deliberated judgement, in which the judgement layer works the scenario space in advance, deposits assessed contingencies with explicit tripwires, and thereby moves human judgement earlier in time rather than out of the loop, avoiding the out-of-the-loop degradation that Endsley (1995) documented without pretending that a deliberative cycle can run at machine tempo. The OODA lineage (Osinga, 2007) survives inside this architecture as the special theory of the time-critical lane — valid there, and disastrous as a general theory, for the reasons Table 1 recorded.

Figure 6. The decision architecture: legitimate configurations by tempo and equivocality

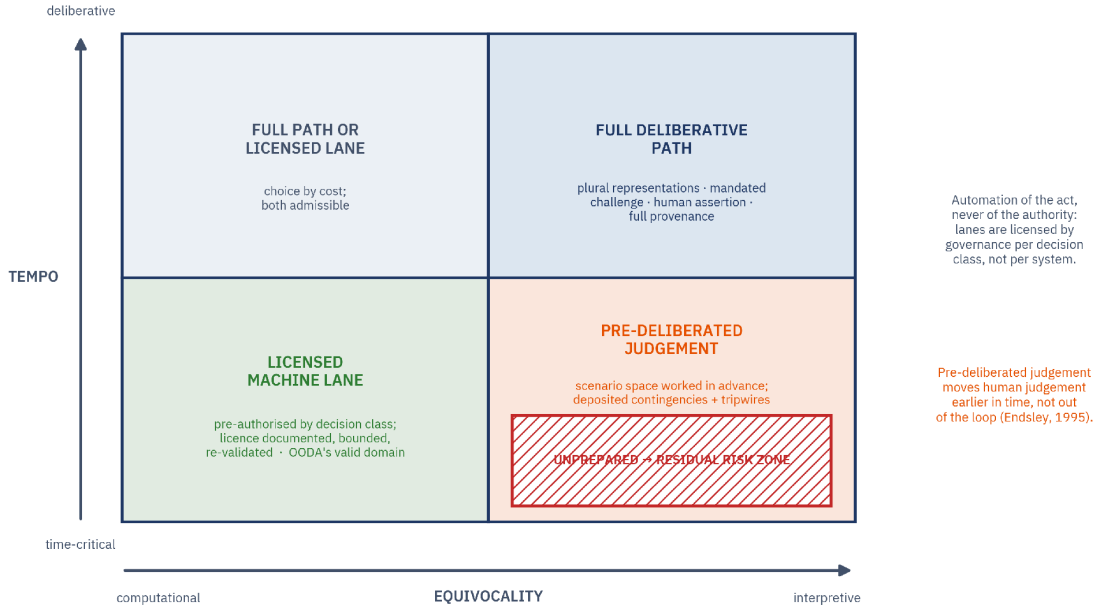

deliberative
FULL PATH OR LICENSED LANE
choice by cost; both admissible
FULL DELIBERATIVE PATH
plural representations · mandated challenge · human assertion · full provenance
TEMPO
LICENSED MACHINE LANE
pre-authorised by decision class; licence documented, bounded, re-validated · OODA's valid domain
PRE-DELIBERATED JUDGEMENT
scenario space worked in advance; deposited contingencies + tripwires
UNPREPARED → RESIDUAL RISK ZONE
Automation of the act, never of the authority: lanes are licensed by governance per decision class, not per system.
Pre-deliberated judgement moves human judgement earlier in time, not out of the loop (Endsley, 1995).
time-critical
computational
EQUIVOCALITY
interpretive

**Figure 6. The decision architecture: legitimate configurations by tempo and equivocality**

## 4.7 Causal logic

The propositions stated in the research design can now be read as the causal skeleton of the framework rather than as free-standing conjectures. P1 (layer separation improves reliability) follows from the assertion monopoly and the failure analysis of Table 10: collapsing sensing or fusion into judgement launders unexamined machine product into assertion, and collapsing judgement into fusion removes the agent capable of adversarial interpretation. P2 (value of machine output is conditional on explainability) follows from Definition 4 and the trust literature: without decomposability there is nothing for calibration to grip, and the system defaults to one of the two miscalibrations (Dietvorst et al., 2015; Logg et al., 2019). P3 (absent cognitive defence, open-source inputs progressively corrupt) follows from the attack-surface analysis and the economics of machine-scaled influence operations (Goldstein et al., 2023): contamination compounds through L1 and model retraining unless L4 interrupts it. P4 (comparative-advantage allocation outperforms) is the direct consequence of Definition 2 applied to Table 12. P5 (design-time governance outperforms post-hoc oversight) follows from the displacement argument of Section 2.12: where the consequential choices are architectural, only architectural governance reaches them. P6 (feedback quality dominates component quality) follows from Section 4.5: a system with mediocre components and functioning loops corrects itself; a system with excellent components and corrupted loops compounds its errors. Each proposition names the variables an empirical test would require, and Section 10 (future research) specifies candidate designs.

**Table 14. Nomological structure of the propositions: antecedents, mechanisms, moderators and outcomes**

| P | Antecedent condition | Mechanism | Moderator(s) | Outcome construct | Candidate indicator |
|---|---|---|---|---|---|
| P1 | Layer-separating vs layer-collapsing architecture | Unexamined machine product barred from assertion; adversarial interpretation retained | Adversarial intensity | Unrecognised assessment error rate | Blind-scored error audits over matched problem sets |
| P2 | Decomposability level of machine products | Interrogability enables trust calibration | Error salience (mediator: trust calibration) | Reliance–reliability coupling | Correlation of analyst reliance with measured system reliability |
| P3 | Presence or absence of a cognitive-defence layer | L4 interrupts contamination compounding through L1 and retraining | Machine-mediation share; provenance coverage | Input-stream corruption gradient | Canary persistence and detection latency in instrumented pipelines |
| P4 | Allocation rule: adversarially assessed advantage vs alternatives | Exploitable machine failure modes withheld from adversary reach | Equivocality; injection intensity | Team analytic performance | Accuracy and manipulation-detection rates under injection |
| P5 | Governance timing: design-time vs post-hoc | Oversight reaches the locus of consequential choice | Design-time decision share | Undetected failure rate | Incident and audit-finding rates across matched organisations |
| P6 | Loop quality relative to component quality | Error correction compounds across cycles; corrupted loops amplify error | Volatility; adversary adaptation rate | Error-correction capacity | Recovery curves in simulation and field proxies |

# 5. Methodology

## 5.1 Research design

This is a conceptual paper, and the choice requires defence rather than apology, since conceptual research operates under quality criteria of its own that are neither weaker nor vaguer than empirical ones — merely different (Gilson & Goldberg, 2015). In Jaakkola's (2020) typology of conceptual designs, the study combines two: theory synthesis, in which fragmented literatures are integrated under a new conceptual lens, and model development, in which constructs and their relationships are specified to the point of testability. The combination is dictated by the problem. Section 3 established that the deficiency in the field is integrative — ten literatures, each partial, none connected — and an integrative deficiency cannot be repaired by another empirical study inside any single silo; it requires the construction of the connecting theory first. MacInnis (2011) distinguishes conceptual goals of identifying, revising, delineating, summarising, differentiating and integrating; the present study's goal is integration in her strict sense, the creation of a framework within which previously unconnected findings become jointly interpretable.

The design also honours the field's own methodological tradition. Intelligence studies' foundational contributions — Kent (1949), Herman (1996), Omand (2010) — are works of conceptual synthesis grounded in documented practice, not hypothesis tests; and the discipline's recent self-criticism (Marrin, 2011, 2017) demands not less theory but better-specified theory. The study accordingly holds itself to Whetten's (1989) canonical requirements for theoretical contribution: explicit statement of the what (the constructs — Definitions 1 to 7), the how (their relations — the layers, interfaces and loops of Sections 4.2 to 4.6), the why (the causal logic — Section 4.7) and the who-where-when (the boundary conditions — Sections 4.1 and 9). Cornelissen's (2017) warning that conceptual papers fail when they blur propositions, process models and typologies into an undifferentiated mass is answered by separation of duties: the framework supplies a process architecture, Table 12 and the decision typology of Section 4.6 supply typologies, and Propositions 1 to 6 supply falsifiable relational claims, each held distinct.

## 5.2 Literature identification and review procedure

The review in Section 2 follows the logic, though not the full protocol, of systematic reviewing (Tranfield, Denyer & Smart, 2003), adapted for a theory-building purpose in the manner Webster and Watson (2002) describe: concept-centric rather than author-centric, organised by theme, and pursued to saturation within each theme rather than to exhaustive enumeration. Ten thematic domains were defined ex ante from the research questions — intelligence process theory, strategic and military intelligence, OSINT, AI and machine learning, decision analytics, knowledge management and fusion, sensemaking and bias, structured techniques, explainable AI and ethics, human–machine teaming and governance including cognitive security. Within each domain, searching combined database queries (Scopus, Web of Science, IEEE Xplore, Google Scholar) with backward and forward citation snowballing from anchor works, prioritising peer-reviewed publications from 2018 onwards while retaining foundational works whose claims remain load-bearing. Inclusion required

that a work either state a theoretical position on the domain's core question or report evidence bearing directly on one; commentary without either was excluded. Two honest limitations of the procedure are recorded here rather than buried in Section 9: no formal inter-coder reliability was possible in a sole-authored conceptual study, and the ten-domain structure, while derived from the research questions, inevitably shapes what the review can see — a hermeneutic circularity that conceptual research can discipline but not escape (Weick, 1989).

## 5.3 Comparative synthesis and framework construction

Synthesis proceeded in three deliberate passes. The first pass produced, for each domain, the position map recorded in the review's tables: who claims what, against whom, with what acknowledged limitation. The second pass read the maps against one another, asking a single disciplined question of each pairing — what does literature A assume that literature B has shown to be false or contested? — from which the five gaps of Section 3 were derived; the gap matrix (Table 9) is the audit trail of that pass. The third pass was constructive and followed Weick's (1989) account of theorising as disciplined imagination: candidate architectures were generated and subjected to thought-trial against the assembled findings, retained where they absorbed the findings without contradiction and discarded where they merely renamed a problem. Two examples of discarded candidates illustrate the discipline. An early three-layer variant merged governance into judgement; it fell to the design-time displacement argument (Section 2.12), which shows that a governance function located inside the judgement layer arrives after the consequential choices have been made. A variant treating cognitive defence as a fifth sequential stage fell to the circularity of Gap 3, which entails that defence must operate concurrently at every layer or protect nothing. Construct definitions were then tightened to Suddaby's (2010) standard — definitional clarity, scope conditions, semantic relationship to neighbouring constructs — which is why Definitions 2 to 7 specify their contrast cases and their limits rather than only their content.

## 5.4 Quality criteria and methodological limitations

The study can be judged on four criteria appropriate to its design. Construct clarity: are the definitions precise, bounded and distinguishable from neighbours (Suddaby, 2010)? Integrative adequacy: does the framework absorb the reviewed findings without contradiction, and would specialists in each of the ten domains recognise their literature fairly stated? Generative capacity: does the framework produce claims that could be false — Propositions 1 to 6 — and research it did not begin from (Section 10)? Practical interpretability: can practitioners locate their own enterprise within the architecture and derive design consequences (Section 8)? What the methodology cannot deliver is equally explicit: no conceptual study validates its own framework, and nothing in this article should be read as evidence that DHI-conforming enterprises outperform others. That is the empirical programme the propositions define, not a result the paper claims — a boundary Section 9 restates and respects.

# 6. Discussion: The Framework Against Its Alternatives

A new framework earns its place by comparison, not assertion. This section sets Intelligence Theory 2026 against the six families of models a reviewer would rightly demand it confront — the traditional cycle, the target-centric process model, the DIKW hierarchy, the OODA loop, evidence-based decision models and hybrid intelligence systems — asking in each case what the incumbent explains, what it cannot, what the framework adds, and, no less deliberately, what the incumbent still does better. Table 16 summarises the comparison, and Figure 9 renders it graphically; Table 17 records what the framework consciously inherits from and rejects in each predecessor, since a theory that cannot name its debts has not understood its own construction.

## 6.1 Against the traditional intelligence cycle

The cycle's virtues are real and should be conceded at the outset: it is teachable, it names functions that any intelligence enterprise must perform, and it has coordinated professional vocabulary across agencies and decades (Lowenthal, 2022). Its failures are the ones Section 2.1 assembled — descriptive falsity (Hulnick, 2006), silence about agency, sequential fiction — and the framework's response is surgical rather than demolitionist (Figure 7): the functions survive as content of the layers; the sequence is replaced by continuous operation and four specified loops; the silence about agency is replaced by the allocation theory of Table 11. What the framework adds that no repair of the cycle could supply is the defensive dimension: the cycle has no location at which manipulation of its own inputs can even be represented, because its ontology contains collection targets but not adversaries of the process itself. What the cycle retains over the framework is economy. Five words teach the cycle; the framework requires seven definitions and four loops, and Section 6.7 treats that cost honestly.

Figure 7. The classical intelligence cycle and its DHI redistribution

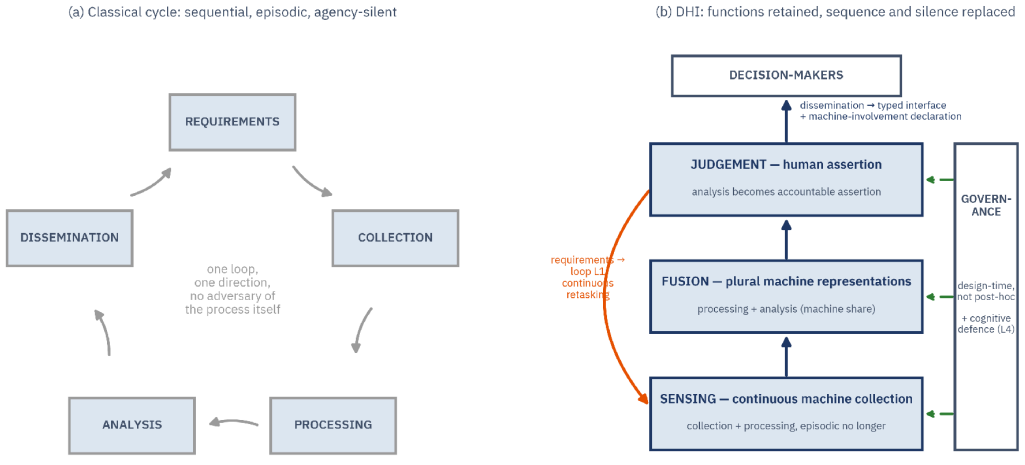

(a) Classical cycle: sequential, episodic, agency-silent
REQUIREMENTS
COLLECTION
PROCESSING
ANALYSIS
DISSEMINATION
one loop, one direction, no adversary of the process itself
(b) DHI: functions retained, sequence and silence replaced
DECISION-MAKERS
dissemination → typed interface + machine-involvement declaration
JUDGEMENT — human assertion
analysis becomes accountable assertion
FUSION — plural machine representations
processing + analysis (machine share)
SENSING — continuous machine collection
collection + processing, episodic no longer
requirements → loop L1 continuous retasking
GOVERN-ANCE
design-time, not post-hoc
+ cognitive defence (L4)

**Figure 7. The classical intelligence cycle and its DHI redistribution**

## 6.2 Against the target-centric process model

Clark's (2019) model anticipated two of the framework's commitments — the abandonment of sequence and the co-construction of a shared representation by collectors, analysts and consumers — and the fusion layer's plural representations are a direct descendant of his shared target model. The divergence is at the point Section 2.1 identified as the model's underspecification: validation. In a target-centric enterprise, who certifies the shared model, against what, and with what protection from an adversary who understands that corrupting one collaborative representation corrupts everything downstream? The framework's answers are structural: plurality (competing representations rather than one), the assertion monopoly (the shared model informs but never is the assessment) and cognitive defence at the representation's inputs. The target-centric model is, in the framework's terms, a fusion-layer theory generalised beyond its layer; it remains the best available account of that layer's collaborative character.

## 6.3 Against the DIKW hierarchy

DIKW (Ackoff, 1989) persists in doctrine because it offers what layered architectures always offer — an intuition of ascent from raw to refined — and Rowley (2007) showed long ago that the intuition dissolves under pressure: the tiers lack defensible boundaries and the ascent lacks a mechanism. The framework's layers are not DIKW renamed, and the difference is worth stating exactly. DIKW's strata are epistemic grades of content; the framework's strata are functional allocations of work, distinguished by agent class, output type and obligation, with the transformation mechanisms — fusion operations, mandated challenge, assertion — specified rather than gestured at. The deeper break is directional: DIKW is a one-way ladder, while the framework's loops make descent (retasking, downgrading, withdrawal) as theoretically important as ascent. What DIKW retains is its service as a reminder that data is not knowledge — a reminder the sensing layer's no-assertion rule institutionalises rather than merely repeats.

## 6.4 Against the OODA loop

Boyd's construct, reconstructed by Osinga (2007), is the framework's most instructive rival because it is the one most often invoked to justify exactly the design the framework forbids: acceleration of the whole enterprise to machine tempo. The framework's treatment, developed in Section 4.6, is containment rather than rejection. Boyd’s OODA logic remains perfectly sound strictly within the time-critical-computational quadrant, and the licensed machine lane implements it. However, when generalized to the entire enterprise, OODA logic becomes a liability—a systematic sacrifice of verification on the altar of speed, which under adversarial information conditions is a gift to the adversary — an opponent who cannot beat the enterprise's judgement can still beat its clock, feeding it material whose refutation takes longer than the decision window. Pre-deliberated judgement is the framework's answer to Boyd's genuine insight (tempo matters) without Boyd's cost (judgement evacuated): the orientation work is done ahead of the clock, not inside it. OODA survives, in short, as a special theory whose domain the decision architecture bounds.

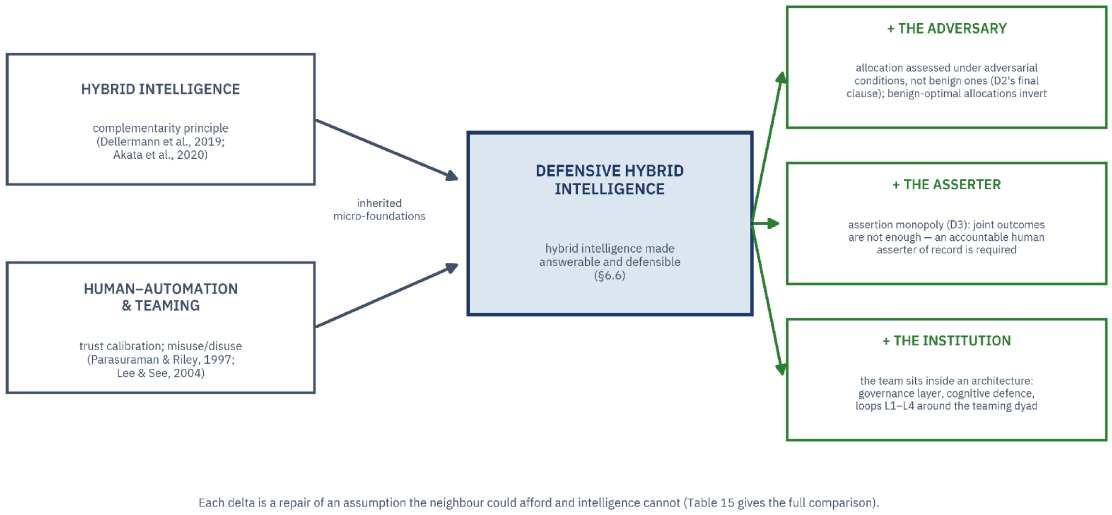
## 6.5 Against evidence-based decision models

The decision-science programme reviewed in Section 2.5 — calibration, scoring, probabilistic discipline (Dhami et al., 2015; Tetlock & Gardner, 2015; Mandel & Barnes, 2014) — is not a rival the framework opposes but an engine it installs: loop L2 and the judgement layer's tradecraft rules are that programme given an architectural home. The comparison is nonetheless substantive, because the programme's standalone form carries two assumptions the framework corrects. The first is the honest-evidence assumption already documented: scoring disciplines judgement given the inputs, and does nothing about poisoned inputs — calibration against a manipulated world produces confident, well-calibrated error. The second is institutional: the programme addresses analysts and their techniques, and Section 2.12 showed the consequential choices migrating to design time, where analyst-level reform cannot reach. Evidence-based models make the judgement layer better; they have no theory of the other three layers, which is why they belong inside a framework rather than in place of one.

## 6.6 Against hybrid intelligence systems

The nearest neighbour deserves the most care. Hybrid intelligence (Dellermann et al., 2019; Akata et al., 2020) shares the framework's central commitment — complementary human–machine configuration — and supplies, as Section 2.11 acknowledged, its micro-foundations. Three additions (Figure 8) mark the distance between a hybrid intelligence system and Defensive Hybrid Intelligence, and each is a repair of an assumption the hybrid literature could afford and intelligence cannot. First, the adversary: hybrid intelligence optimises complementarity in benign environments; DHI allocates by comparative advantage assessed under adversarial conditions (Definition 2's final clause), which changes allocations — tasks machines perform best in the laboratory move back to humans where the machine's failure mode is the one the adversary can trigger. Second, assertion: the hybrid literature speaks of joint outcomes; intelligence requires an accountable asserter, and the assertion monopoly is a constraint on hybridity that generic hybrid theory has no reason to impose. Third, the enterprise: hybrid intelligence theorises the team; DHI theorises the institution around the team — governance, defence, loops — without which teamed excellence sits inside an undefended and unaccountable pipeline. DHI is, compactly, hybrid intelligence made answerable and made defensible.

Figure 8. What Defensive Hybrid Intelligence adds to its nearest neighbours


HYBRID INTELLIGENCE
complementarity principle (Dellermann et al., 2019; Akata et al., 2020)
HUMAN–AUTOMATION & TEAMING
trust calibration; misuse/disuse (Parasuraman & Riley, 1997; Lee & See, 2004)
inherited micro-foundations
DEFENSIVE HYBRID INTELLIGENCE
hybrid intelligence made answerable and defensible (§6.6)
+ THE ADVERSARY
allocation assessed under adversarial conditions, not benign ones (D2's final clause); benign-optimal allocations invert
+ THE ASSERTER
assertion monopoly (D3): joint outcomes are not enough — an accountable human asserter of record is required
+ THE INSTITUTION
the team sits inside an architecture: governance layer, cognitive defence, loops L1–L4 around the teaming dyad

Each delta is a repair of an assumption the neighbour could afford and intelligence cannot (Table 15 gives the full comparison).

**Figure 8. What Defensive Hybrid Intelligence adds to its nearest neighbours**

## Table 15. Differentiation from the nearest neighbours: hybrid intelligence and human–automation research

| Dimension | Hybrid intelligence (Dellermann et al., 2019; Akata et al., 2020) | Human–automation & teaming (Parasuraman & Riley, 1997; Lee & See, 2004) | Defensive Hybrid Intelligence (this framework) |
|---|---|---|---|
| Environment assumption | Benign; errors statistical | Benign; reliability varies, nobody manipulates it | Adversarial; inputs, models and cognition under deliberate attack (D5) |
| Allocation criterion | Complementarity of capabilities | Function allocation by workload and reliability | Comparative epistemic advantage assessed under adversarial conditions (D2) |
| Output ontology | Joint human–machine outcome | Operator action supported by automation | Typed products; observations and representations distinct from assertions |
| Accountable asserter | Unspecified / diffuse | Operator, implicitly | Judgement layer by design — assertion monopoly (D3) |
| Unit of theory | The task and the team | The operator–system dyad | The enterprise: layers, interfaces, loops, governance |
| Defence function | Absent | Absent | Standing cognitive defence over the attack surface (D6, loop L4) |
| Status of explainability | Design desideratum | Reliability information for calibration | Load-bearing admissibility requirement with a security function (D4) |
| Governance location | External to the theory | External to the theory | Design-time layer inside the architecture (§4.2.4) |

## 6.7 What the framework costs

Symmetry requires the ledger's other side. The framework is more expensive than every model it confronts, in three currencies. Pedagogically, it cannot be taught in a diagram and five words; adoption requires the kind of doctrinal investment the cycle never demanded. Operationally, its obligations — provenance immutability, mandated challenge, plural representations, standing defence — consume analyst time and machine capacity that a pipeline architecture would spend on throughput, and the framework's wager that the spending buys reliability is exactly that, a wager, pending the empirical programme. Organisationally, the assertion monopoly and design-time governance redistribute authority — away from tool owners and towards the judgement and governance functions — and redistributions of authority are resisted for reasons that have nothing to do with theory (Jensen et al., 2020). A candid comparison therefore ends unheroically: for a small enterprise in a benign information environment performing routine monitoring, the classical models remain cheaper and adequate; the framework's claim is conditional, holding where machine analysis is substantive, open sources dominant and the environment contested — which is, however, precisely the condition Section 1 argued is becoming general.

**Table 16. Systematic comparison of Intelligence Theory 2026 with incumbent models**

| Dimension | Traditional cycle | Target-centric | DIKW | OODA | Intelligence Theory 2026 (DHI) |
|---|---|---|---|---|---|
| Process form | Sequential ring | Networked co-construction | One-way ascent | Rapid closed loop | Layered, continuous, four specified feedback loops |
| Agency of analysis | Human (implicit) | Human collaborative | Unspecified | Human commander | Allocated by comparative epistemic advantage (Table 11) |
| Machine role | Absent / tool | Supporting the shared model | Absent | Accelerant | Constitutive at sensing and fusion; barred from assertion |
| Adversarial information environment | Not representable | Not addressed | Not addressed | Tempo contest only | Structural: epistemic attack surface + cognitive defence + L4 |
| Explainability | Not addressed | Implicit in collaboration | Not addressed | Not addressed | Load-bearing interface obligation (Definition 4) |
| Governance | External oversight | Not addressed | Not addressed | Not addressed | Design-time layer woven through architecture |

| Dimension | Traditional cycle | Target-centric | DIKW | OODA | Intelligence Theory 2026 (DHI) |
|---|---|---|---|---|---|
| Accountability for judgement | Analyst (by custom) | Diffuse | Unspecified | Commander | Assertion monopoly (Definition 3) |
| Adaptive mechanism | Implicit feedback | Model updating | None | Orientation | Loops L1–L4 with specified corruptions |
| Principal cost | Descriptive falsity | Validation gap | Mechanism gap | Verification sacrificed | Doctrinal, operational and organisational expense (§6.7) |

**Table 17. What the framework inherits and what it rejects, by predecessor**

| Predecessor | Inherited into the framework | Rejected, with grounds |
|---|---|---|
| Traditional cycle | The functional inventory: tasking, collection, processing, analysis, dissemination survive as layer content | Sequence and stage ontology; silence on agency; inability to represent attacks on the process itself |
| Target-centric model | Co-constructed representations; abandonment of linearity (fusion layer lineage) | Single shared model (replaced by mandated plurality); unspecified validation authority |
| DIKW hierarchy | The raw/refined distinction, institutionalised as the sensing layer's no-assertion rule | Epistemic strata as architecture; one-way ascent; absent transformation mechanisms |
| OODA loop | Tempo as a real variable; contained as the licensed machine lane's logic | Tempo as the master value; enterprise-wide acceleration; verification subordinated to speed |
| Evidence-based decision models | Scoring, calibration and probabilistic discipline, installed as loop L2 and judgement tradecraft | Standalone sufficiency; honest-evidence assumption; analyst-level scope |
| Hybrid intelligence | Complementarity principle; trust-calibration micro-foundations | Benign-environment allocation; outcome without accountable asserter; team without institution |

Figure 9. Framework comparison: incumbent models and DHI across nine dimensions

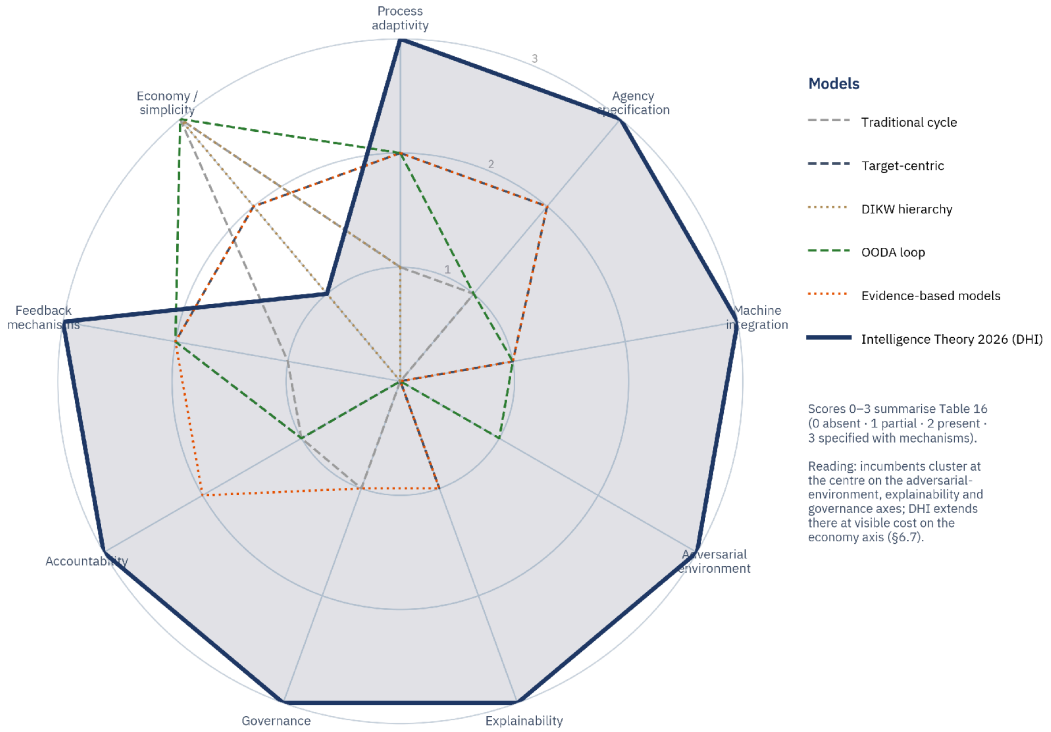



**Figure 9. Framework comparison: incumbent models and DHI across nine dimensions**

# 7. Theoretical Contributions

It is intellectual good practice to explicitly close the loops opened at the outset. Table 18 does not merely summarize; it directly confronts the initial research questions with the theoretical architecture developed throughout the study, mapping each to the sections that answer it.

**Table 18. Closure of the research questions**

| RQ | Answered in | Answer in one sentence |
|---|---|---|
| RQ1 | §2.1; §6.1–6.4 | The canonical models fail at agency, sequence and the representability of attacks on the intelligence process itself. |
| RQ2 | §4.3 | Seven constructs are necessary and jointly sufficient: DHI, comparative epistemic advantage, the assertion monopoly, the explainability obligation, the epistemic attack surface, cognitive defence and the trust-calibration interface (Definitions 1–7; Table 11). |
| RQ3 | §2.3; §4.2.1 | OSINT is repositioned from a collection discipline among others to the primary, contested substrate of the sensing layer, ingested under permanent input defence. |
| RQ4 | §4.4 | Explainability functions as the precondition of trust calibration and as a security instrument, which converts it from a model property into an admissibility requirement at the machine-to-human interface. |
| RQ5 | §4.2.4 | Cognitive defence is integrated as a standing governance function operating concurrently at every layer through loop L4, not as a separate discipline or sequential stage. |
| RQ6 | §4.2.4; §8 | Accountability requires design-time architecture — provenance, model documentation, machine-involvement declaration — because machine mediation relocates consequential choices beyond the reach of retrospective oversight. |
| RQ7 | §4.5 | Four specified loops — tasking, calibration, model behaviour and defence — carry learning and error correction, and each has a named corruption mode that governance must monitor (Table 13). |

Four contributions are claimed, stated here in the descending order of their generality and with their beneficiaries named, since a contribution that benefits nobody in particular usually benefits nobody at all.

**First, epistemic security enters intelligence theory as a first-order property.** Existing theory treats deception as a problem for analysis and disinformation as a problem for society; neither treatment locates the protection of the enterprise's own knowledge production inside the theory of that production. The DHI construct (Definition 1) and the epistemic attack surface (Definition 5) perform that relocation, and the framework demonstrates why it is not optional: under conditions where the collection substrate, the analytic models and the consumers' cognition are simultaneously exposed to the same adversarial activity, an architecture without structural defence does not produce degraded intelligence — it produces confident intelligence whose degradation is invisible to it. The immediate beneficiary is intelligence studies itself, which acquires a vocabulary for a class of failure its canonical models cannot express; the FIMI and cognitive-security communities benefit reciprocally, gaining a theoretical seat inside the discipline whose taxonomies they have so far supplied from outside (EEAS, 2023; Pamment, 2020).

**Second, the human–machine boundary is theorised rather than assumed.** The literatures reviewed in Sections 2.4, 2.5 and 2.11 agree that allocation matters and decline to specify it; the framework specifies it through three connected constructs — comparative epistemic advantage assessed under adversarial conditions (Definition 2), the assertion monopoly (Definition 3) and the allocation mapping of Table 12. The theoretical move is the adversarial clause: allocation criteria developed in benign environments (Jarrahi, 2018; Dellermann et al., 2019) can recommend delegating to machines precisely those tasks whose machine failure modes an adversary can trigger, and the framework shows that allocation theory for intelligence must therefore be a branch of adversarial reasoning, not of ergonomics. Beneficiaries: hybrid intelligence research, which gains a hard case that stress-tests its complementarity principle, and intelligence organisations, which gain a principled alternative to allocation by procurement momentum.

**Third, explainability is converted from a technical property into a tradecraft requirement,** with a specification (Definition 4) that is simultaneously more demanding than the XAI mainstream — decomposability to the weakest dependency, because explanation functions as a security instrument under manipulation — and more achievable, because it demands product-level accountability rather than global model transparency (Section 4.4). The contribution is a genuine bridge: intelligence studies obtains a requirement it can write into doctrine and procurement; XAI research obtains a user community whose explanation needs are concrete, adversarial and professionally sceptical — precisely the conditions Miller (2019) argued the field lacks. The unresolved question the bridge exposes — whether foundation models can meet the obligation at judgement-critical junctions or must be displaced there by interpretable architectures (Rudin, 2019) — is itself a contribution, since it converts a diffuse anxiety into a decidable research question (Section 10).

**Fourth, governance moves from the outside of the theory to the inside of the architecture.** The displacement argument of Section 2.12 — that machine mediation relocates consequential choices to design time, beyond the reach of retrospective oversight — has a cognate in the algorithmic-accountability literature, where Ananny and Crawford (2018) trace the failure of transparency ideals against systems whose consequential choices are architectural; its statement as a problem for intelligence theory is, to the author's knowledge, new here, and the governance layer (Section 4.2.4) is its constructive answer: accountability, model governance, cognitive defence and legitimacy management as standing functions woven through production rather than reviews applied to its products. Oversight scholarship gains a specification of what design-time oversight would have to reach; ethics-of-intelligence scholarship gains what Sections 2.10 identified as missing — a location inside the architecture for principles that otherwise decorate it (Floridi et al., 2018; Blanchard & Taddeo, 2023).

Cutting across the four is the integrative contribution recorded in Table 9 and executed in Section 4: ten literatures that did not previously constrain one another now do. How this changes intelligence theory can be stated compactly. The discipline's central question shifts from the classical form — how should information become judgement? — to the form the

framework makes tractable: how should an enterprise of humans and machines produce accountable judgement from a contested information environment it cannot assume to be honest? The classical question survives inside the new one, as the judgement layer; the new question could not even be posed inside the old vocabulary.

## 8. Practical Implications

Because the framework is a theory of organisation rather than a technology prescription, its practical implications take the same form for every audience — locate your enterprise within the architecture, identify which layers, interfaces and loops are absent or collapsed, and treat the absences as risk registers — while differing sharply in emphasis. Eight audiences are addressed; Table 19 summarises.

For government intelligence agencies, the nearest-term implications are the two monopolies: assertion and provenance. Agencies adopting language models for drafting face, today, the exact condition the assertion monopoly forbids — fluent machine text entering assessment streams without mandated challenge — and the framework supplies the doctrinal language for prohibiting it without prohibiting the tools. The provenance obligation implies procurement criteria: systems that cannot expose retrieval traceability and confidence derivation fail the explainability obligation regardless of accuracy, a criterion procurement officers can apply without adjudicating model internals. For military organisations, the decision architecture of Section 4.6 is the operative contribution: the licensed machine lane gives commanders a governed alternative to the false choice between full automation and full deliberation, and pre-deliberated judgement gives doctrine writers a named, specifiable method for keeping human judgement inside time-critical engagements by moving it earlier — with the tripwire designs it requires identified as a research need (Section 10) rather than asserted as solved.

For corporate and competitive intelligence functions, the framework travels with one substitution — regulatory and reputational exposure standing in for national oversight — and one warning that travels undiluted: commercial OSINT pipelines feeding commercial language models are exposed to input manipulation exactly as state pipelines are, and a competitor or short-seller occupies the adversary position in the attack-surface analysis as naturally as a foreign service does. For cybersecurity operations, the mapping is nearly one-to-one — threat intelligence is already a sensing-fusion-judgement stack — and the framework's marginal contribution is loop L4: treating detection of manipulation as a signal that automatically re-weights upstream sources, rather than as a ticket closed upon attribution. Critical infrastructure operators inherit the time-critical-interpretive quadrant as their characteristic condition, which makes pre-deliberated judgement — worked scenario spaces, deposited contingencies, explicit tripwires — their principal import from the framework.

For policymakers, two implications outrank the rest. The oversight implication: legislative and judicial mechanisms built to review products and warrants will not reach design-time choices, and statutes governing intelligence AI should mandate the auditability the governance layer specifies — provenance retention, model documentation,

machine-involvement declaration in assessments — because these are the preconditions of any future review. The EU's Artificial Intelligence Act supplies the template and the warning at once: its design-time obligations are the right species of instrument, and its national-security exemptions leave the intelligence enterprise largely outside them — a gap legislatures will have to close deliberately rather than by drift. The procurement implication: explainability requirements should be written at product level, where they are specifiable and testable, not at model level, where they are neither. For AI developers, the framework is effectively a requirements document from a demanding customer: typed outputs that distinguish observation from assertion; confidence propagated from source reliability rather than asserted post hoc; decomposability on demand within operational time; behavioural monitoring hooks as first-class interfaces (Rahwan et al., 2019). Developers who read Table 10's failure-mode column as a test plan will build what the judgement layer can actually use. For universities and professional education, the implication is curricular: intelligence education that teaches the cycle as architecture is now teaching a special case, and the analytically decisive skills the framework identifies — interrogating machine products, reasoning about manipulation of one's own inputs, calibrated probability under scoring — are teachable and largely untaught; the framework supplies a syllabus skeleton, and the EPIC ten-sprint structure of İnal (2026) illustrates one full curriculum built on these commitments.

**Table 19. Practical implications by audience**

| Audience | Primary import from the framework | Nearest-term action | Characteristic risk if ignored |
|---|---|---|---|
| Government agencies | Assertion monopoly; explainability obligation as procurement criterion | Mandated-challenge doctrine for machine-drafted analysis | Fluent unaccountable text colonises assessment streams |
| Military organisations | Decision architecture: licensed lanes and pre-deliberated judgement | Type decision classes; license automation by class, not by system | Enterprise-wide OODA acceleration; verification sacrificed |
| Corporate intelligence | Attack-surface analysis with commercial adversaries | Contamination monitoring on OSINT-to-LLM pipelines | Confident assessments built on seeded sources |
| Cybersecurity operations | Loop L4: detections re-weight upstream sources automatically | Wire threat-intel detections into source reliability scoring | Manipulation detected but not propagated; repeated compromise |
| Critical infrastructure | Pre-deliberated judgement for time-critical-interpretive events | Scenario spaces with deposited contingencies and tripwires | Improvised judgement at machine tempo during incidents |
| Policymakers | Design-time oversight; product-level explainability mandates | Statutory auditability: provenance, documentation, declarations | Oversight audits the shadow while pipelines decide |
| AI developers | Table 10 failure modes as requirements; typed, decomposable outputs | Build provenance, confidence propagation and monitoring hooks | Technically excellent systems inadmissible to judgement |
| Universities | Curriculum: machine interrogation, input defence, calibrated judgement | Replace cycle-as-architecture teaching with layer-and-loop literacy | Graduates trained for an enterprise that no longer exists |

# 9. Limitations

Six limitations bound what this article has shown, and the first is the one that governs all readings of it: the framework is proposed, not validated. Nothing here demonstrates that DHI-conforming enterprises outperform alternatives; Propositions 1 to 6 are the claims on which the framework stakes itself, and until the empirical programme of Section 10 has been run, adoption arguments rest on the plausibility of the derivations in Section 4.7, not on evidence of effect. Second, the supporting literature carries a systematic bias the synthesis inherits: it is predominantly Anglophone, Western and state-centric, and an architecture distilled from it may encode institutional assumptions — legislative oversight traditions, particular civil-military relations, resource levels — that do not travel; the framework's claim to describe functions rather than organisations is a hypothesis about its own portability, not a demonstration of it. Third, the constructs face genuine operationalisation difficulty: 'decomposability to the weakest dependency within operational time' and 'cognitive defence maturity' name measurable-in-principle properties for which no instruments yet exist, and until they do, the propositions are testable in design but not in practice — a gap Section 10 places first on the agenda.

Fourth, the technological referents will age asymmetrically: the framework was constructed to be architecture-level precisely so that its claims survive model turnover, but its illustrative reasoning is anchored in the foundation-model era (Bommasani et al., 2021), and capabilities that collapsed the collection-analysis boundary once may collapse other boundaries the architecture currently treats as stable — the assertion monopoly, in particular, is a normative commitment whose practical enforceability against increasingly agentic systems — whose governance is only beginning to be theorised (Chan et al., 2024; Gabriel et al., 2024) — is an open question, not a settled one. Fifth, the framework's obligations create their own failure mode, which Section 2.8 supplies the precedent for: structured analytic techniques decayed into ritual compliance, and mandated challenge, provenance discipline and trust-calibration interfaces can decay the same way — the framework specifies architecture, and architecture cannot compel the analytic culture that animates it (Johnston, 2005). Sixth, scope: Section 6.7 stated the boundary condition and it bears repeating as a limitation — for small enterprises, benign environments and routine monitoring, the framework is over-engineered, and its costs are certain while its benefits are, at present, propositional.

**Table 20. Boundary conditions: where the framework applies and where it does not**

| Dimension | Framework applies | Framework does not apply (or reduces) | Rationale |
|---|---|---|---|
| Machine involvement | Substantive machine participation in sensing, fusion or drafting | Fully human enterprises — DHI reduces to classical tradecraft plus loop L2 | The hybrid half is vacuous without machine analytic labour |
| Information environment | Contested environments with sustained FIMI exposure | Verifiably benign environments — classical models are cheaper and adequate (§6.7) | The defensive half prices in an adversary; absent one, its costs buy nothing |
| Enterprise scale | Multi-analyst enterprises able to separate functions | Micro-teams where one person necessarily spans all layers | Layer separation and the assertion monopoly presuppose role differentiation |
| Decision tempo | All four quadrants of §4.6, via the three licensed configurations | Pure machine-speed closed-loop control with no assertion event | Where nothing is asserted, the monopoly has no object; such systems fall to control engineering |
| Institutional authority | Actors able to enforce design-time governance on their machine estate | Consumers of opaque third-party analysis with no procurement leverage | Design-time governance requires authority over design |
| Temporal scope | The foundation-model capability regime (epochal naming, §4.1) | Capability regimes in which agentic systems defeat enforceable assertion (§9) | The framework's normative commitments assume enforceability |

# 10. Future Research

The agenda divides into three programmes — testing the propositions, building the instruments, and extending the boundaries — comprising seventeen directions (Figure 10). Table 22 maps the first programme onto the propositions it serves.

Figure 10. The research programme: three tracks, seventeen directions, six propositions

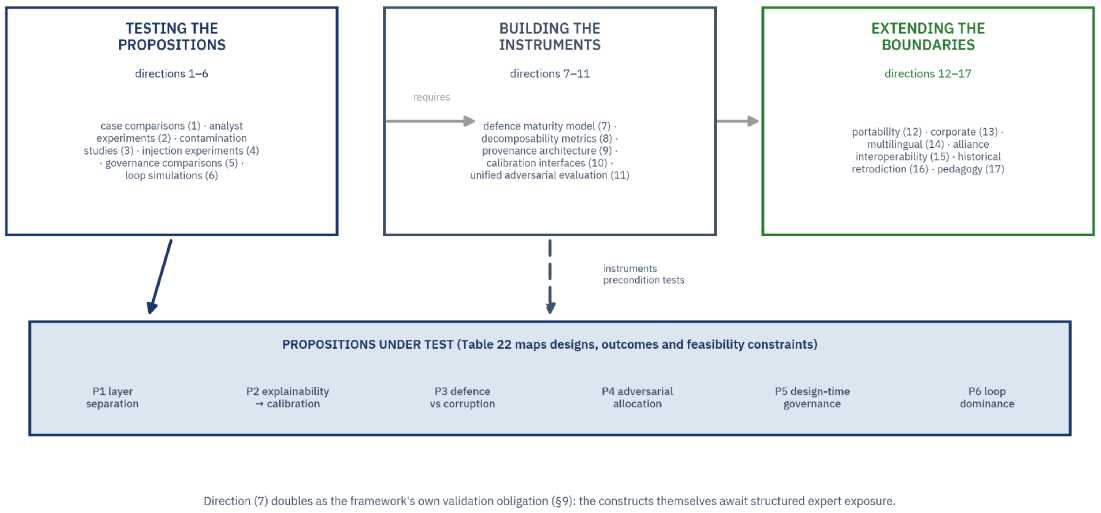

TESTING THE PROPOSITIONS
directions 1–6
case comparisons (1) · analyst experiments (2) · contamination studies (3) · injection experiments (4) · governance comparisons (5) · loop simulations (6)
BUILDING THE INSTRUMENTS
directions 7–11
requires
defence maturity model (7) · decomposability metrics (8) · provenance architecture (9) · calibration interfaces (10) · unified adversarial evaluation (11)
EXTENDING THE BOUNDARIES
directions 12–17
portability (12) · corporate (13) · multilingual (14) · alliance interoperability (15) · historical retrodiction (16) · pedagogy (17)
instruments precondition tests
PROPOSITIONS UNDER TEST (Table 22 maps designs, outcomes and feasibility constraints)
P1 layer separation
P2 explainability → calibration
P3 defence vs corruption
P4 adversarial allocation
P5 design-time governance
P6 loop dominance
Direction (7) doubles as the framework's own validation obligation (§9): the constructs themselves await structured expert exposure.

**Figure 10. The research programme: three tracks, seventeen directions, six propositions**

## 10.1 Testing the propositions

(1) Comparative case studies of intelligence enterprises whose architectures separate versus collapse the sensing-fusion-judgement functions, using assessment reliability over matched problem sets as the outcome, would test P1; declassified post-mortems and commercial intelligence firms offer feasible sites. (2) Controlled experiments with professional analysts, varying the decomposability of machine products while holding accuracy constant and measuring trust calibration (Lee & See, 2004) and error detection, would test P2 — extending the Vogel et al. (2021) findings from observation to manipulation. (3) Longitudinal contamination studies — instrumented OSINT pipelines with known-provenance canaries and FIMI-taxonomy monitoring (EEAS, 2023) — would measure the corruption gradient P3 asserts, and would constitute, incidentally, the first quantitative estimate of a quantity the entire disinformation debate currently lacks. (4) Task-allocation experiments in which human-machine teams face problem sets with and without adversarial injection would test P4's central claim that benign-environment allocations invert under attack. (5) Comparative governance research across agencies and firms that differ in design-time versus post-hoc oversight, with incident and failure rates as outcomes, would test P5 — difficult access, but the corporate AI-governance field offers proxy sites. (6) Multi-agent simulation of layered architectures, varying component quality against loop quality, would test P6 cheaply before

any field study, and would identify the loop-corruption dynamics of Table 13 that field studies should instrument.

## 10.2 Building the instruments

(7) Measurement development for the framework's constructs — a cognitive-defence maturity model with behaviourally anchored levels, and decomposability metrics such as time-to-weakest-dependency under analyst interrogation — is the precondition of the entire first programme. (8) Provenance architecture research for LLM-mediated pipelines: what technical form must immutable provenance take when transformations include generative summarisation, and what compression of provenance is admissible before the explainability obligation fails? (9) Interface research on trust-calibration displays for professional sceptics, a user class the human-AI interaction literature (Amershi et al., 2019) has barely studied. (10) Design research on pre-deliberated judgement: methods for scenario-space coverage, tripwire specification and contingency staleness — the decision architecture's least developed component. (11) Adversarial evaluation methodology that unifies the currently separate practices of adversarial machine learning and FIMI red-teaming into a single model-defence discipline for loop L4.

**Table 21. Operationalisation of the constructs: candidate indicators and measurement approaches**

| Construct | Candidate indicator(s) | Measurement approach |
|---|---|---|
| D1 DHI conformance | Architecture conformance score across layers, interfaces and loops | Structured audit checklist derived from Table 10 and §4.5; inter-rater reliability across auditors |
| D2 Comparative epistemic advantage | Allocation-rule fidelity; share of tasks allocated against Table 12 with documented rationale | Task-log analysis; deviation register review |
| D3 Assertion monopoly | Share of disseminated assessments with an accountable human asserter of record; incidence of unadopted machine text in products | Product audit with provenance reconstruction |
| D4 Explainability obligation | Time-to-weakest-dependency under analyst interrogation; decomposition depth achieved within the decision-class window | Timed interrogation protocol with professional analysts (direction 2) |
| D5 Epistemic attack surface | Exposure-inventory coverage: fraction of inputs, models and consumer channels with a maintained exposure record | Inventory audit against pipeline architecture |
| D6 Cognitive defence | Maturity level (behaviourally anchored model, direction 7); canary detection rate; mean time from detection to upstream re-weighting | Instrumented-pipeline telemetry (direction 3); red-team exercise scoring |
| D7 Trust calibration interface | Reliance–reliability correlation (Lee & See, 2004); miscalibration asymmetry (bias vs aversion incidence) | Controlled analyst experiments varying reliability display (direction 2) |

## 10.3 Extending the boundaries

(12) Portability studies: application of the framework to small-state, non-Western and non-state intelligence enterprises, testing the state-centric bias named in Section 9 and identifying which functions are scale-invariant. (13) Corporate adaptation: a translated framework for competitive intelligence in which regulatory exposure replaces oversight and market adversaries replace foreign services. (14) Multilingual extension: contamination dynamics and defence taxonomies in non-English information environments, where FIMI activity is substantial and research coverage thin. (15) Alliance interoperability: what must be true of two enterprises' governance layers — provenance standards, declaration formats, defence-signal sharing — for DHI-conforming intelligence to be shareable without laundering one partner's unexamined machine product into another's judgement layer? (16) Historical retrodiction: systematic re-reading of documented intelligence failures through the framework's categories, asking which failures the architecture would have represented that the cycle could not — a discipline-native validation method with a long tradition (Betts, 1978; Dahl, 2013). (17) Pedagogical research: curriculum designs and learning outcomes for layer-and-loop literacy, machine interrogation and calibrated judgement, testing whether the skills Section 8 declared teachable are in fact taught best together or apart.

**Table 22. The empirical programme mapped to the propositions**

| Proposition | Research design (direction number) | Primary outcome variable | Principal feasibility constraint |
|---|---|---|---|
| P1 | Comparative case studies of layer-separating vs layer-collapsing architectures (1) | Assessment reliability over matched problems | Access; architecture classification validity |
| P2 | Analyst experiments varying product decomposability (2) | Trust calibration; error detection rate | Professional-analyst sample recruitment |
| P3 | Instrumented pipelines with canaries and FIMI monitoring (3) | Contamination gradient over time | Ethics of canary injection; attribution noise |
| P4 | Team experiments with and without adversarial injection (4) | Allocation-performance interaction | Constructing realistic injection without deception harms |
| P5 | Comparative governance studies, design-time vs post-hoc (5) | Incident and failure rates | Confounded organisational variables |
| P6 | Multi-agent simulation: loop quality vs component quality (6) | System error correction over cycles | Simulation validity; parameter grounding |

## 11. Conclusion

This article began from a discrepancy that the discipline has registered piecemeal and not yet theorised whole: the canonical models of intelligence production assume scarce information, human analytic monopoly and a broadly honest information environment, and all three assumptions are now false. The response developed here is not a further critique of the intelligence cycle — the literature contains excellent ones — but a successor of comparable generality: Intelligence Theory 2026, organised around Defensive Hybrid Intelligence, in which four functionally specified layers allocate analytic labour by comparative epistemic advantage assessed under adversarial conditions, every machine-to-human interface carries a decomposability obligation strong enough for tradecraft, four feedback loops with named corruption modes make the architecture adaptive, and the protection of the enterprise's own inputs, models and analysts is a standing function rather than an adjacent speciality.

The framework's originality was argued by audit rather than assertion: ten literatures were mapped, their mutual silences recorded in a gap matrix, and each element of the architecture was derived as the repair of a specific, documented absence — with the framework's debts to its predecessors itemised as carefully as its departures, and its costs stated in the same register as its claims. Its testability was built in rather than appended: six propositions carry the causal skeleton, and the research agenda specifies the designs, outcome variables and feasibility constraints their testing requires. What the article deliberately does not claim is validation. The framework stands, at this point, as a disciplined answer to a question the field's inherited vocabulary could not express: how an enterprise of humans and machines should produce accountable judgement from an information environment that is simultaneously its richest source, its models' training corpus and its adversaries' preferred battlefield. If the propositions survive the programme now specified, intelligence theory will have acquired an architecture adequate to its era; if they fail, the failures will be informative in a way the present vocabulary's silences are not. Either outcome is progress, and that — not any single construct — is the standard a successor theory owes the tradition it revises.

# References (consolidated, APA 7)


Ackoff, R. L. (1989). From data to wisdom. Journal of Applied Systems Analysis, 16(1), 3–9.

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138–52160.

Akata, Z., Balliet, D., de Rijke, M., et al. (2020). A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. Computer, 53(8), 18–28.

Akhgar, B., Bayerl, P. S., & Sampson, F. (Eds.). (2016). Open source intelligence investigation: From strategy to implementation. Springer.

Alavi, M., & Leidner, D. E. (2001). Knowledge management and knowledge management systems: Conceptual foundations and research issues. MIS Quarterly, 25(1), 107–136.

Amershi, S., Weld, D., Vorvoreanu, M., et al. (2019). Guidelines for human-AI interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1–13). ACM.

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society, 20(3), 973–989.

Anderljung, M., Barnhart, J., Korinek, A., et al. (2023). Frontier AI regulation: Managing emerging risks to public safety. arXiv. https://arxiv.org/abs/2307.03718

Anthropic. (2024). Introducing the Model Context Protocol. Anthropic. https://www.anthropic.com/news/model-context-protocol

Bai, Y., Kadavath, S., Kundu, S., et al. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv. https://arxiv.org/abs/2212.08073

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115.

Bellaby, R. W. (2014). The ethics of intelligence: A new framework. Routledge.

Bengio, Y., Mindermann, S., Privitera, D., et al. (2025). International AI safety report (arXiv:2501.17805). AI Safety Summit / UK Department for Science, Innovation and Technology.

Benkler, Y., Faris, R., & Roberts, H. (2018). Network propaganda: Manipulation, disinformation, and radicalization in American politics. Oxford University Press.

Betts, R. K. (1978). Analysis, war, and decision: Why intelligence failures are inevitable. World Politics, 31(1), 61–89.

Blanchard, A., & Taddeo, M. (2023). The ethics of artificial intelligence for intelligence analysis: A review of the key challenges with recommendations. Digital Society, 2(1), Article 12. https://doi.org/10.1007/s44206-023-00036-4

Bommasani, R., Hudson, D. A., Adeli, E., et al. (2021). On the opportunities and risks of foundation models. arXiv. https://arxiv.org/abs/2108.07258

Born, H., & Leigh, I. (2005). Making intelligence accountable: Legal standards and best practice for oversight of intelligence agencies. Publishing House of the Parliament of Norway.

Bradshaw, S., & Howard, P. N. (2019). The global disinformation order: 2019 global inventory of organised social media manipulation. Oxford Internet Institute.

Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Buchanan, B. (2020). The AI triad and what it means for national security strategy. Center for Security and Emerging Technology.

Bullwinkel, B., Minnich, A., Chawla, S., et al. (2025). Lessons from red teaming 100 generative AI products. arXiv. https://arxiv.org/abs/2501.07238

Carlini, N., Jagielski, M., Choquette-Choo, C. A., Paleka, D., Pearce, W., Anderson, H., Terzis, A., Thomas, K., & Tramèr, F. (2024). Poisoning web-scale training datasets is practical. In 2024 IEEE Symposium on Security and Privacy (SP) (pp. 407–425). IEEE.

Chan, A., Ezell, C., Kaufmann, M., et al. (2024). Visibility into AI agents. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24). ACM.

Chang, W., Berdini, E., Mandel, D. R., & Tetlock, P. E. (2018). Restructuring structured analytic techniques in intelligence. Intelligence and National Security, 33(3), 337–356.

Choo, C. W. (1996). The knowing organization: How organizations use information to construct meaning, create knowledge and make decisions. International Journal of Information Management, 16(5), 329–340.

Clark, R. M. (2019). Intelligence analysis: A target-centric approach (6th ed.). CQ Press.

Coalition for Content Provenance and Authenticity. (2023). Content Credentials: C2PA technical specification. C2PA.

Cornelissen, J. (2017). Editor's comments: Developing propositions, a process model, or a typology? Addressing the challenges of writing theory without a boilerplate. Academy of Management Review, 42(1), 1–9.

Coulthart, S. J. (2017). An evidence-based evaluation of 12 core structured analytic techniques. International Journal of Intelligence and CounterIntelligence, 30(2), 368–391.

Dahl, E. J. (2013). Intelligence and surprise attack: Failure and success from Pearl Harbor to 9/11 and beyond. Georgetown University Press.

Davies, P. H. J. (2004). Intelligence culture and intelligence failure in Britain and the United States. Cambridge Review of International Affairs, 17(3), 495–520.

Debenedetti, E., Zhang, J., Balunović, M., Beurer-Kellner, L., Fischer, M., & Tramèr, F. (2024). AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. In Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track.

Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid intelligence. Business & Information Systems Engineering, 61(5), 637–643.

Department for Science, Innovation and Technology. (2024). Introduction to AI assurance. UK Government.

Dhami, M. K., Mandel, D. R., Mellers, B. A., & Tetlock, P. E. (2015). Improving intelligence analysis with decision science. Perspectives on Psychological Science, 10(6), 753–757.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1), 114–126.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv. https://arxiv.org/abs/1702.08608

du Cluzel, F. (2021). Cognitive warfare (Innovation Hub study, June–November 2020). NATO Allied Command Transformation, Innovation Hub.

Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of big data: Evolution, challenges and research agenda. International Journal of Information Management, 48, 63–71.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. Human Factors, 37(1), 32–64.

European Commission. (2025). General-purpose AI code of practice. European Commission.

European External Action Service. (2023). 1st EEAS report on foreign information manipulation and interference threats. EEAS.

European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L series, 12 July 2024.

Feffer, M., Sinha, A., Deng, W. H., Lipton, Z. C., & Heidari, H. (2024). Red-teaming for generative AI: Silver bullet or security theater? Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 7(1), 421–437.

Floridi, L., Cowls, J., Beltrametti, M., et al. (2018). AI4People — An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689–707.

Freelon, D., & Wells, C. (2020). Disinformation as political communication. Political Communication, 37(2), 145–156.

Gabriel, I., Manzini, A., Keeling, G., et al. (2024). The ethics of advanced AI assistants. arXiv. https://arxiv.org/abs/2404.16244

Ganguli, D., Lovitt, L., Kernion, J., et al. (2022). Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. arXiv. https://arxiv.org/abs/2209.07858

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé, H., III, & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86–92.

Gentry, J. A., & Gordon, J. S. (2019). Strategic warning intelligence: History, challenges, and prospects. Georgetown University Press.

Ghioni, R., Taddeo, M., & Floridi, L. (2024). Open source intelligence and AI: A systematic review of the GELSI literature. AI & Society, 39(4), 1827–1842. https://doi.org/10.1007/s00146-023-01628-x

Gill, P., & Phythian, M. (2018). Intelligence in an insecure world (3rd ed.). Polity Press.

Gilson, L. L., & Goldberg, C. B. (2015). Editors' comment: So, what is a conceptual paper? Group & Organization Management, 40(2), 127–130.

Glassman, M., & Kang, M. J. (2012). Intelligence in the internet age: The emergence and evolution of Open Source Intelligence (OSINT). Computers in Human Behavior, 28(2), 673–682.

Goldfarb, A., & Lindsay, J. R. (2022). Prediction and judgment: Why artificial intelligence increases the importance of humans in war. International Security, 46(3), 7–50.

Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., & Sedova, K. (2023). Generative language models and automated influence operations: Emerging threats and potential mitigations. arXiv. https://arxiv.org/abs/2301.04246

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt

injection. In Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec '23) (pp. 79–90). ACM.

Gunning, D., & Aha, D. W. (2019). DARPA's explainable artificial intelligence (XAI) program. AI Magazine, 40(2), 44–58.

Hall, D. L., & Llinas, J. (1997). An introduction to multisensor data fusion. Proceedings of the IEEE, 85(1), 6–23.

Herman, M. (1996). Intelligence power in peace and war. Cambridge University Press.

Heuer, R. J. (1999). Psychology of intelligence analysis. Center for the Study of Intelligence, Central Intelligence Agency.

Heuer, R. J., & Pherson, R. H. (2021). Structured analytic techniques for intelligence analysis (3rd ed.). CQ Press.

Horowitz, M. C. (2018). Artificial intelligence, international competition, and the balance of power. Texas National Security Review, 1(3), 37–57.

Hribar, G., Podbregar, I., & Ivanuša, T. (2014). OSINT: A 'grey zone'? International Journal of Intelligence and CounterIntelligence, 27(3), 529–549.

Hulnick, A. S. (2006). What's wrong with the intelligence cycle. Intelligence and National Security, 21(6), 959–979.

İnal, Ç. (2026). Intelligence theory 2026. EVENTICA Publications.

International Organization for Standardization. (2023). ISO/IEC 42001:2023 — Information technology — Artificial intelligence — Management system. ISO.

Jaakkola, E. (2020). Designing conceptual articles: Four approaches. AMS Review, 10(1–2), 18–26.

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. Business Horizons, 61(4), 577–586.

Jensen, B. M., Whyte, C., & Cuomo, S. (2020). Algorithms at war: The promise, peril, and limits of artificial intelligence. International Studies Review, 22(3), 526–550.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399.

Johnston, R. (2005). Analytic culture in the US intelligence community: An ethnographic study. Center for the Study of Intelligence, Central Intelligence Agency.

Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. American Psychologist, 64(6), 515–526.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). Noise: A flaw in human judgment. Little, Brown Spark.

Kendall, W. (1949). The function of intelligence. World Politics, 1(4), 542–552.

Kent, S. (1949). Strategic intelligence for American world policy. Princeton University Press.

Khaleghi, B., Khamis, A., Karray, F. O., & Razavi, S. N. (2013). Multisensor data fusion: A review of the state-of-the-art. Information Fusion, 14(1), 28–44.

Klein, G. (1998). Sources of power: How people make decisions. MIT Press.

Klein, G., Moon, B., & Hoffman, R. R. (2006). Making sense of sensemaking 1: Alternative perspectives. IEEE Intelligent Systems, 21(4), 70–73.

Lahneman, W. J. (2010). The need for a new intelligence paradigm. International Journal of Intelligence and CounterIntelligence, 23(2), 201–225.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50–80.

Lipton, Z. C. (2018). The mythos of model interpretability. Communications of the ACM, 61(10), 36–43.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes, 151, 90–103.

Lowenthal, M. M. (2022). Intelligence: From secrets to policy (9th ed.). CQ Press.

MacInnis, D. J. (2011). A framework for conceptual contributions in marketing. Journal of Marketing, 75(4), 136–154.

Mandel, D. R., & Barnes, A. (2014). Accuracy of forecasts in strategic intelligence. Proceedings of the National Academy of Sciences, 111(30), 10984–10989.

Marrin, S. (2011). Improving intelligence analysis: Bridging the gap between scholarship and practice. Routledge.

Marrin, S. (2017). Understanding and improving intelligence analysis by learning from other disciplines. Intelligence and National Security, 32(5), 539–547.

McGregor, S. (2021). Preventing repeated real world AI failures by cataloging incidents: The AI Incident Database. Proceedings of the AAAI Conference on Artificial Intelligence, 35(17), 15458–15463.

Mercado, S. C. (2004). Sailing the sea of OSINT in the information age. Studies in Intelligence, 48(3). https://www.cia.gov/resources/csi/studies-in-intelligence/archives/vol-48-no-3/

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1–38.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19) (pp. 220–229). ACM.

National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). U.S. Department of Commerce.

National Institute of Standards and Technology. (2024). Artificial intelligence risk management framework: Generative artificial intelligence profile (NIST AI 600-1). U.S. Department of Commerce.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. Review of General Psychology, 2(2), 175–220.

Nonaka, I., & Takeuchi, H. (1995). The knowledge-creating company. Oxford University Press.

Omand, D. (2010). Securing the state. Hurst.

Omand, D., & Phythian, M. (2018). Principled spying: The ethics of secret intelligence. Oxford University Press.

OpenAI. (2023). GPT-4 system card. OpenAI.

OpenAI. (2024). Model Spec. OpenAI. https://model-spec.openai.com

Osinga, F. P. B. (2007). Science, strategy and war: The strategic theory of John Boyd. Routledge.

Pamment, J. (2020). The EU's role in fighting disinformation: Taking back the initiative. Carnegie Endowment for International Peace.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39(2), 230–253.

Pastor-Galindo, J., Nespoli, P., Gómez Mármol, F., & Martínez Pérez, G. (2020). The not yet exploited goldmine of OSINT: Opportunities, open challenges and future trends. IEEE Access, 8, 10282–10304.

Payne, K. (2021). I, warbot: The dawn of artificially intelligent conflict. Hurst.

Perez, E., Huang, S., Song, F., et al. (2022). Red teaming language models with language models. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 3419–3448). ACL.

Phythian, M. (Ed.). (2013). Understanding the intelligence cycle. Routledge.

Pirolli, P., & Card, S. (2005). The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In Proceedings of the International Conference on Intelligence Analysis. MITRE.

Rahwan, I., Cebrian, M., Obradovich, N., et al. (2019). Machine behaviour. Nature, 568(7753), 477–486.

Rid, T. (2020). Active measures: The secret history of disinformation and political warfare. Farrar, Straus and Giroux.

Rønn, K. V., & Høffding, S. (2013). The epistemic status of intelligence: An epistemological contribution to the understanding of intelligence. Intelligence and National Security, 28(5), 694–716.

Roozenbeek, J., van der Linden, S., Goldberg, B., Rathje, S., & Lewandowsky, S. (2022). Psychological inoculation improves resilience against misinformation on social media. Science Advances, 8(34), Article eabo6254.

Rowley, J. (2007). The wisdom hierarchy: Representations of the DIKW hierarchy. Journal of Information Science, 33(2), 163–180.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215.

Shevlane, T., Farquhar, S., Garfinkel, B., et al. (2023). Model evaluation for extreme risks. arXiv. https://arxiv.org/abs/2305.15324

Shrestha, Y. R., Ben-Menahem, S. M., & von Krogh, G. (2019). Organizational decision-making structures in the age of artificial intelligence. California Management Review, 61(4), 66–83.

Simon, H. A. (1955). A behavioral model of rational choice. The Quarterly Journal of Economics, 69(1), 99–118.

Starbird, K., Arif, A., & Wilson, T. (2019). Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), Article 127.

Steele, R. D. (2007). Open source intelligence. In L. K. Johnson (Ed.), Handbook of intelligence studies (pp. 129–147). Routledge.

Steinberg, A. N., Bowman, C. L., & White, F. E. (1999). Revisions to the JDL data fusion model. In Proceedings of SPIE 3719, Sensor Fusion: Architectures, Algorithms, and Applications III (pp. 430–441). SPIE.

Suddaby, R. (2010). Editor's comments: Construct clarity in theories of management and organization. Academy of Management Review, 35(3), 346–357.

Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. Science, 361(6404), 751–752.

Tetlock, P. E., & Gardner, D. (2015). Superforecasting: The art and science of prediction. Crown.

Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. British Journal of Management, 14(3), 207–222.

Treverton, G. F. (2009). Intelligence for an age of terror. Cambridge University Press.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185(4157), 1124–1131.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

Verdoliva, L. (2020). Media forensics and deepfakes: An overview. IEEE Journal of Selected Topics in Signal Processing, 14(5), 910–932.

Vogel, K. M., Reid, G., Kampe, C., & Jones, P. (2021). The impact of AI on intelligence analysis: Tackling issues of collaboration, algorithmic transparency, accountability, and management. Intelligence and National Security, 36(6), 827–848.

Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policymaking. Council of Europe.

Warner, M. (2002). Wanted: A definition of 'intelligence'. Studies in Intelligence, 46(3), 15–22.

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. MIS Quarterly, 26(2), xiii–xxiii.

Weick, K. E. (1989). Theory construction as disciplined imagination. Academy of Management Review, 14(4), 516–531.

Weick, K. E. (1995). Sensemaking in organizations. Sage.

Whetten, D. A. (1989). What constitutes a theoretical contribution? Academy of Management Review, 14(4), 490–495.

Williams, H. J., & Blum, I. (2018). Defining second generation open source intelligence (OSINT) for the defense enterprise. RAND Corporation.

Zegart, A. B. (2022). Spies, lies, and algorithms: The history and future of American intelligence. Princeton University Press.

Zhan, Q., Liang, Z., Ying, Z., & Kang, D. (2024). InjecAgent: Benchmarking indirect prompt injections in tool-integrated large language model agents. In Findings of the Association for Computational Linguistics: ACL 2024. ACL.

Zou, W., Geng, R., Wang, B., & Jia, J. (2025). PoisonedRAG: Knowledge corruption attacks to retrieval-augmented generation of large language models. In Proceedings of the 34th USENIX Security Symposium (pp. 3827–3844). USENIX Association.

# Appendix A: Literature Review Summary Tables

**Table 1. Canonical models of the intelligence process: assumptions, criticisms and status under contemporary conditions**

| Model | Principal sources | Core assumption | Principal criticism | Status under 2026 conditions |
|---|---|---|---|---|
| Traditional intelligence cycle | Kent (1949); Lowenthal (2022) | Sequential stages; separation of collection, analysis and consumption | Descriptively false; stages operate in parallel (Hulnick, 2006) | Breaks down when collection is continuous and machine-mediated |
| Target-centric model | Clark (2019) | Shared model of the target maintained collaboratively by all participants | Underspecifies who validates the shared model and how conflict is resolved | Compatible with AI participation but silent on machine epistemology and governance |
| OODA loop | Boyd, in Osinga (2007) | Advantage accrues to the actor who observes, orients, decides and acts fastest | Speed emphasis suits tactical contest, not deliberative strategic assessment | Acceleration by AI amplifies the model's neglect of verification and reflection |
| DIKW hierarchy | Ackoff (1989); Rowley (2007) | Value ascends linearly from data through information and knowledge to wisdom | Linearity and stage boundaries are conceptually unstable (Rowley, 2007) | Machine systems traverse the lower tiers instantly, collapsing the hierarchy's explanatory value |
| Intelligence as epistemic practice | Rønn & Høffding (2013); Betts (1978) | Intelligence is judgement under uncertainty; failure is structurally conditioned | Offers diagnosis rather than an operational architecture | Correct in orientation but pre-dates hybrid human–machine judgement |

**Table 2. Conceptual positions on OSINT: claims and unresolved issues**

| Position | Representative works | Core claim | Principal unresolved issue |
|---|---|---|---|
| Advocacy / substitution | Steele (2007); Mercado (2004) | Open sources can satisfy most intelligence requirements at lower cost and risk than secret collection | Retains classical epistemology; offers no account of verification at scale |
| Transformation | Glassman & Kang (2012) | Internet-era OSINT changes the nature of intelligence problem-solving, not merely its inputs | Underspecifies institutional design implications |
| Second-generation OSINT | Williams & Blum (2018) | Social media and big-data analytics dissolve the collection–analysis boundary | Framework is descriptive periodisation rather than explanatory theory |
| Computational OSINT | Pastor-Galindo et al. (2020); Akhgar et al. (2016) | OSINT is an engineering problem of acquisition, resolution and credibility assessment | Credibility assessment under adversarial manipulation remains unsolved |

| Position | Representative works | Core claim | Principal unresolved issue |
|---|---|---|---|
| Critical / governance | Hribar et al. (2014); Ghioni et al. (2024) | OSINT operates in a legal-ethical grey zone; governance lags capability | No integration of governance requirements into intelligence process theory |

**Table 3. Competing positions on the role of AI in intelligence and strategy**

| School | Representative works | Core claim | Principal limitation |
|---|---|---|---|
| Augmentation | Vogel et al. (2021); Zegart (2022) | AI automates analytic drudgery; humans retain judgement; transparency and accountability are the binding constraints | Assumes a benign information environment; does not theorise model poisoning or adversarial inputs |
| Substitution-sceptical | Goldfarb & Lindsay (2022) | Machine prediction complements and raises the value of human judgement | Economic framing abstracts from analytic tradecraft and institutional detail |
| Transformation | Horowitz (2018); Payne (2021) | AI reshapes strategic competition and the character of conflict; diffusion and adaptation decide advantage | Operates above the level of intelligence process design |
| Organisational limits | Jensen, Whyte & Cuomo (2020) | Institutional absorption, not algorithmic capability, is the binding constraint | Diagnostic; offers no architecture for successful absorption |
| Threat-side | Goldstein et al. (2023); Buchanan (2020) | Generative models industrialise influence operations and create data-dependency vulnerabilities | Not connected to intelligence production theory; defence treated as external to analysis |

**Table 4. Decision-making research traditions and their bearing on intelligence analysis**

| Tradition | Representative works | Core claim | Limitation for intelligence theory |
|---|---|---|---|
| Bounded rationality | Simon (1955) | Decision-makers satisfice under cognitive and informational limits | Says little about how organisations should structure analytic support |
| Heuristics and biases | Tversky & Kahneman (1974); Kahneman (2011) | Judgement departs from normative standards in systematic, predictable ways | Debiasing prescriptions weakly validated in operational settings |
| Naturalistic decision-making | Klein (1998); Kahneman & Klein (2009) | Expert intuition is recognition-based and environment-dependent | Strategic intelligence is a low-validity environment where intuition is least reliable |
| Forecasting and decision science | Tetlock & Gardner (2015); Mandel & Barnes (2014); Dhami et al. (2015) | Judgemental accuracy is measurable, trainable and manageable | Assumes an honest evidence base; adversarial manipulation untheorised |
| Organisational AI decision-making | Shrestha et al. (2019); Duan et al. (2019); Jarrahi (2018) | Human–machine configurations trade off speed, interpretability and accountability | Delegation boundary unresolved; no security or adversarial dimension |

**Table 5. Information fusion levels (JDL, revised) mapped to intelligence functions**

| JDL level | Engineering description | Intelligence equivalent | Open problem under adversarial conditions |
|---|---|---|---|
| Level 0–1: signal and object refinement | Detection, alignment, entity estimation and tracking | Processing and exploitation; collation of raw take | Sensor spoofing; fabricated digital artefacts; synthetic media |
| Level 2: situation assessment | Relations among entities; contextual aggregation | All-source analysis of current situations | Poisoned open-source context; coordinated inauthentic corroboration |
| Level 3: impact assessment | Projection of consequences and threat intent | Estimative and warning intelligence | Adversary shaping of the evidence base used for projection |
| Level 4: process refinement | Sensor management; resource reallocation | Collection management driven by analytic gaps | Feedback loops themselves become an attack surface |

**Table 6. The structured analytic techniques debate**

| Position | Representative works | Core claim | Weakness identified by critics |
|---|---|---|---|
| Codification | Heuer & Pherson (2021); Heuer (1999) | Externalised, structured reasoning mitigates cognitive bias and enables review | Adopted on face validity; empirical support thin |
| Evidence-based scepticism | Coulthart (2017) | Validation studies are scarce and results mixed; effectiveness unproven | Offers evaluation, not an alternative method |
| Decision-scientific reform | Chang et al. (2018); Dhami et al. (2015) | Techniques should be rebuilt around probabilistic updating and measured accuracy | Institutional adoption barriers; cultural resistance to scoring |
| Professionalisation critique | Marrin (2011, 2017) | Analysis must become a self-testing discipline drawing on other fields | Programmatic; implementation pathway unspecified |

**Table 7. Positions within explainable AI and their implications for analytic tradecraft**

| Position | Representative works | Core claim | Implication for intelligence analysis |
|---|---|---|---|
| Post-hoc explanation | Adadi & Berrada (2018); Barredo Arrieta et al. (2020) | Opaque models can be explained after the fact by attribution and surrogate methods | Explanations may not reflect actual model computation; tradecraft traceability not guaranteed |
| Social-science critique | Miller (2019) | Explanations must be contrastive, selective and audience-appropriate | Analyst-facing explanation must support interrogation, not mere display |
| Interpretable by design | Rudin (2019) | High-stakes domains require inherently interpretable models, not explained black boxes | Judgement-critical pipeline stages may need interpretable architectures despite accuracy costs |

| Position | Representative works | Core claim | Implication for intelligence analysis |
|---|---|---|---|
| Definitional rigour | Doshi-Velez & Kim (2017); Lipton (2018) | Interpretability is undefined and conflates distinct properties | Procurement and doctrine cannot specify requirements the field cannot define |
| Human-performance evaluation | Gunning & Aha (2019) | Explanation quality is measured by user performance and calibrated trust | Trust calibration, not raw accuracy, is the correct evaluation target for analytic AI |

**Table 8. Human–automation and hybrid intelligence findings relevant to analytic architecture**

| Finding | Representative works | Content | Architectural implication |
|---|---|---|---|
| Automation misuse and disuse | Parasuraman & Riley (1997) | Operators over- and under-rely on automation in predictable patterns | Human oversight cannot be assumed effective; it must be designed and measured |
| Trust calibration | Lee & See (2004) | The objective is appropriate trust matched to actual system reliability | Interfaces must convey reliability information, not only outputs |
| Algorithm aversion | Dietvorst et al. (2015) | Seen errors cause abandonment of superior algorithms | Visible error handling and recovery protocols required to sustain use |
| Algorithm appreciation | Logg et al. (2019) | Advice framed as algorithmic is over-weighted by non-experts | Machine-drafted analysis risks unearned authority; provenance labelling essential |
| Hybrid intelligence | Dellermann et al. (2019); Akata et al. (2020) | Complementary human–machine configurations outperform either alone | Task allocation by comparative epistemic advantage is a design principle |
| Machine behaviour | Rahwan et al. (2019) | AI systems must be studied empirically as behaving agents | Deployed analytic models require continuous behavioural monitoring |

# Figure Captions

Ten figures accompany the article, each reproduced in the text at its point of reference. The captions are listed below.

Figure 1 (Section 3) draws the open information environment as a contested common substrate feeding the enterprise's sensing layer, its models' training corpora and its consumers' cognition simultaneously, with FIMI injection and the Gap 3 circularity marked.

Figure 2 (§4.1) renders the epistemic attack surface: the three targets, their characteristic attack vectors, and the three cognitive-defence functions beneath them.

Figure 3 (§4.2) is the architecture diagram: three horizontal layers with governance as a vertical spine, typed products ascending, loops L1–L4 descending in contrasting weight, and attack-surface markers at their points of exposure.

Figure 4 (§4.2.4) details cognitive defence: the three defence functions with their instruments and their automatic propagation paths into each layer.

Figure 5 (§4.3) maps the seven constructs and their relations, the hybrid half above and the defensive half below, joined at the trust-calibration interface.

Figure 6 (§4.6) plots the decision architecture on tempo and equivocality, with the three licensed configurations and the residual risk zone marked.

Figure 7 (§6.1) contrasts the classical five-stage cycle with its DHI redistribution in two panels.
Figure 8 (§6.6) shows what the framework adds to hybrid intelligence and human–automation research: the adversary, the asserter and the institution.
Figure 9 (Section 6) renders Table 16 as a nine-axis comparison profile, the incumbents clustered at the centre on the adversarial, explainability and governance axes and the framework extended there at visible cost on the economy axis.
Figure 10 (Section 10) lays out the research programme: three tracks, seventeen directions and their dependency on the propositions under test.